DOCTORAL THESIS

Quality Assessment of Enhanced Underwater Images with Convolutional Neural Networks

Muhammad Irshad

Brasília, November 25, 2022

UNIVERSITY OF BRASILIA

FACULTY OF TECHNOLOGY

UNIVERSITY OF BRASILIA
Faculty of Technology
Department of Electrical Engineering

DOCTORAL THESIS

Quality Assessment of Enhanced Underwater Images with Convolutional Neural Networks

Muhammad Irshad

Doctoral Thesis submitted for the partial requirement
for the degree of Doctorate in Electrical Engineering

Examination Board

Prof. Dr. Mylene Christine Queiroz de Farias, ENE/UnB _____
Supervisor

Prof. Dr. José Gabriel Rodriguez Carneiro Gomes, UFRJ _____
External Examiner

Prof. Dr. Wamberto José Lira de Queiroz, UFCG _____
External Examiner

Prof. Dr. João Luiz Carvalho, ENE/UnB _____
Internal Examiner

FICHA CATALOGRÁFICA

MUHAMMAD IRSHAD
Quality Assessment of Enhanced Underwater Images with Convolutional Neural Networks [Distrito Federal] 2022.
xvi, 82 p. (PGEA .TD 193/22), 210 x 297 mm (ENE/FT/UnB, Doctorate, Electrical Engineering, 2022).
Doctoral Thesis - University of Brasília, Faculty of Technology.
Department of Electrical Engineering

1. Image Quality Assessment
2. Perceptual Quality Assessment
3. Convolutional Neural Networks
4. Underwater image Enhancement
I. ENE/FT/UnB
II. Title (series)

BIBLIOGRAPHIC REFERENCE

I. MUHAMMAD (2022). *Quality Assessment of Enhanced Underwater Images with Convolutional Neural Networks*. Doctoral Thesis, Department of Electrical Engineering, University of Brasília, Brasília, DF, 82 p. (PGEA .TD 193/22)

ASSIGNMENT OF RIGHTS

AUTOR: Muhammad Irshad
TÍTULO: Quality Assessment of Enhanced Underwater Images with Convolutional Neural Networks.
GRAU: Doctorate in Electrical Engineering ANO: 2022

Muhammad Irshad
Department of Electrical Engineering - ENE - FT
University of Brasília (UnB)
Campus Darcy Ribeiro
CEP: 70919-970 - Brasília - DF - Brasil

I dedicate this work to my beloved wife Sana Alamgeer.

Agradecimentos

I am truly grateful to my advisor Prof. Dr. Mylene C. Q. Farias for her consistent guidance, support, cooperation, and inspiration throughout my Ph.D. journey. My sincere appreciation and gratitude also go to the senior team of the Group of Digital Signal Processing (GPDS) laboratory, especially Dr. Alessandro Silva, Dr. Pedro Garcia, Dr. Max Vizcarra, Dr. Kerlla Luz, Dr. Helard Becerra, Dr. Jonathan Lima, Dr. Gustavo Sandri, and Dr. Rafael Diniz for my research work.

I am obliged and grateful to my wife, Dr. Sana Alamgeer, for always believing and supporting me in completing this work.

I am grateful to all my colleagues in the Laboratory of the Group of Digital Signal Processing (GPDS), specifically Dário Morais, Vinícius Oliveira, Henrique Garcia, Lucas dos Santos, André Henrique, Thayane Viana, and Priscila Andrade for being so kind and cooperative and for making me feel at home.

I am also very thankful to my parents and family members, who have always believed in me and supported me.

Last but not least, I hereby express my sincere gratitude to the University of Brasília for giving me the opportunity to do my doctorate while exploring the Brazilian culture. I also thank the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES) and the Fundação de Apoio a Pesquisa do Distrito Federal (FAPDF) for their financial support, without which my research would not have been possible.

Muhammad Irshad

ABSTRACT

Image enhancement algorithms have the goal of improving the image quality and, therefore, the usefulness of an image for a given task. Although there are several image enhancement algorithms, there is no consensus on how to estimate the performance of these enhancement algorithms. Since the final consumers of the resulting enhanced visual content are human viewers, the performance of these algorithms should take into account the perceived visual quality of the resulting enhanced images. Unfortunately, although in the last decades a lot of progress has been made in the area of image quality assessment, designing metrics to estimate the quality of enhanced and restored images remains a challenge. This is particularly true for underwater image application, where images frequently need to be restored because of the severity of the degradations introduced by the underwater environment. Therefore, there is a great need for quality metrics that can estimate the quality of enhanced and restored images. In this thesis, our goal is to design metrics for this scenario. First, we have designed a quality metric based on texture operators and saliency. Second, we also designed a quality metric based on a deep learning architecture convolutional neural network (CNN). Experimental results on the underwater image database demonstrate that our approaches outperform the state-of-art methods compared. Third, we have developed a new dataset for underwater image quality assessment. Additionally, we also present a psychophysical study based on crowd-sourcing interface, in which we analyze the perceptual quality of images enhanced with several types of enhancement algorithms. In this experiment, we have developed a database that can be used to train image quality metrics, and also can detect both increments and decrements in the perceived quality.

Keywords: Image Quality Assessment, Enhanced Images, Underwater Images, MSLBP, Convolutional Neural Networks.

RESUMO

Título: Avaliação de Qualidade de imagens subaquáticas aprimoradas com redes neurais convolucionais

Os algoritmos de aprimoramento de imagens tem o objetivo de melhorar a qualidade da imagem e, dessa forma, sua utilidade para uma dada tarefa. Embora existam vários algoritmos de aprimoramento de imagem, não há consenso sobre como estimar o desempenho desses algoritmos. Desde que os consumidores finais do conteúdo visual aprimorado são espectadores humanos, o desempenho destes algoritmos devem levar em conta a qualidade visual percebida da imagem resultante. Infelizmente, apesar das últimas década tenha sido feito muito progresso na área de avaliação de qualidade de imagem, projetar métricas para estimar a qualidade das imagens aprimoradas e restauradas continua sendo um desafio. Isto é particularmente verdade para aplicações de imagens subaquáticas, onde as imagens precisam ser restauradas frequentemente devido a severidade das degradações introduzidas pelo ambiente embaixo da água. Desta forma, há uma grande demanda por métricas que possam estimar a qualidade de imagens aprimoradas e restauradas. Nesta tese, nosso objetivo é projetar métricas para este cenário. Primeiro, nós projetamos uma métrica de qualidade baseada em operadores de textura e saliência. Segundo, nós também projetamos uma métrica de qualidade baseada na arquitetura de aprendizado profundo de uma rede neural convolucional. Resultados experimentais em um banco de dados de imagem subaquáticas demonstram que nossas aproximação supera os métodos do estado da arte comparados. Terceiro, desenvolvemos um novo banco de dados para avaliação da qualidade de imagens subaquáticas. Adicionalmente, também apresentamos um estudo psicofísico baseado em interface “crowdsourcing”, no qual analisamos a qualidade percebida de imagens melhoradas com diversos tipos de algoritmos de aprimoramento. Neste experimento, desenvolvemos uma base de dados que pode ser usada para treinar métricas de qualidade que também possam detectar incrementos e decrementos na qualidade percebida.

Palavras-Chave: avaliação de qualidade, imagens aprimoradas, imagens subaquáticas, MSLBP, redes neurais convolucionais.

SUMMARY

List of Figures

LIST OF TABLES

1 INTRODUCTION

1.1 OVERVIEW

In the last 20 years, the field of image quality assessment has experienced a tremendous boom, with a large number of new methods for image quality assessment being developed every year. Digital imaging technologies restructure the way we capture, store, use, and share images. Today, we have the ability to share photos online instantly, send and receive multimedia messages, and stream live video across the globe instantly [6]. There has been a huge increase in the popularity of image-based applications, particularly smartphones, laptops, tablets, and personal computers. The image services offered by these applications have become an essential part of the end-user's life. As a consequence, the popularity of these applications has led to a huge growth in Internet traffic. According to a current report by Cisco™ [7], every second, global internet traffic increases by more than 100,000 gigabytes. In this scenario, image-based applications require quantifying to what extent the content is affected by these operations. For this purpose, applications assess the visual quality of affected (distorted) visual content to ensure that the delivered content meets the requirements of end users. The goal of image quality assessment is to provide quality metrics that can automatically predict perceived image quality.

In summary, image quality assessment (IQA) methods are useful tools for image processing applications and systems. IQA methods can be classified into two categories: subjective and objective IQA methods. Subjective IQA methods gauge the quality of visual content by performing psychophysical experiments in controlled laboratory environments. In psychophysical or subjective experiments, a number of human observers (subjects) analyze the visual quality of the displayed contents. It is common to use a pool of subjects that are diverse (in terms of gender, age, occupation, etc.) and naive to the technologies being tested. The experimental methodology should follow a standardized recommendation, such as Recommendation ITU-R BT.500 [8] and Recommendation ITU-T P.800.1 [9]. In most types of experimental methodologies, subjects are asked to rate the quality or other attribute of the displayed content (e.g., colorfulness, sharpness, noise, etc.). An estimate of quality is given by the mean opinion score (MOS), which is computed by averaging the scores given by all subjects to a visual content test. Figure 1.1 shows a block diagram of a typical subjective IQA experimental methodology. Although subjective methods are considered ground-truth in visual quality assessment, these methods are expensive, time-consuming, and their results are not easy to replicate.

Objective IQA methods use computational models, also known as quality metrics (QM), to estimate the quality of visual content. Objective IQA methods are faster, cheaper, and can be more easily incorporated into image-based applications. In other words, given the limitations of subjective methods, objective methods are often preferred. For this reason, great effort has been made to develop fast and high accuracy quality metrics [6, 10]. Figure 1.2 shows a block dia-

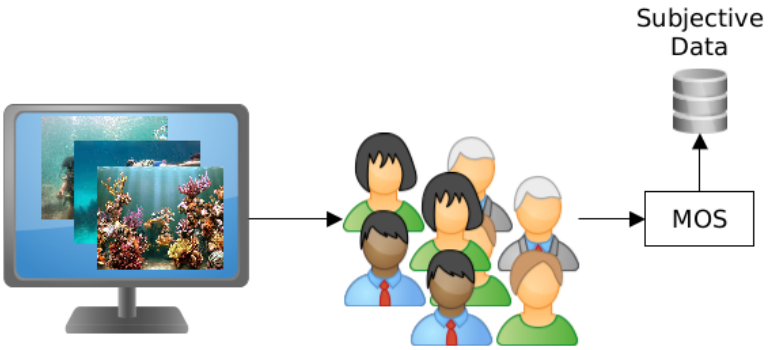

Figure 1.1: Block diagram of a typical Subjective Quality Assessment (SQA) method.

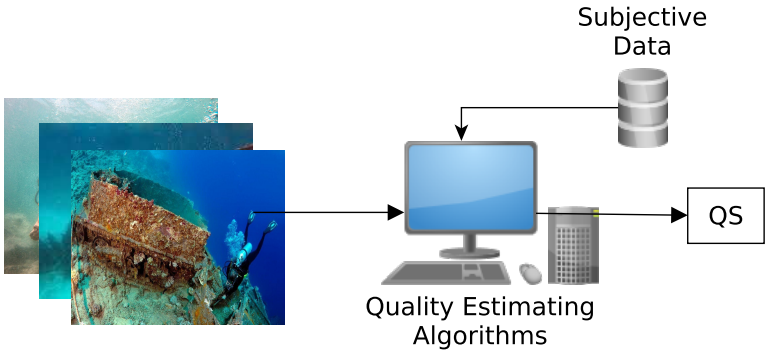

Figure 1.2: Block diagram of a typical Objective Quality Assessment (OQA) method.

gram of a typical objective image quality assessment method. Objective IQA methods are either dedicated quality metrics, which assess a specific type of distortion, or generic quality metrics (also known as general-purpose metrics) that estimate the overall perceived quality. Since the human visual system (HVS) is the ultimate estimator of visual quality, researchers often integrate characteristics of the human visual system in the design of quality metrics. Depending on the amount of reference information used, both dedicated and general-purpose quality metrics are further divided into three types, i.e. full reference (FR) IQA methods, where a reference image is needed to estimate the quality, and no reference (NR) IQA methods, which blindly estimate quality without having access to the reference or pristine image. Reduce reference (RR) IQA methods aim to predict the visual quality of distorted images with only partial information about reference images [11]. The figure 1.3 shows types of objective quality assessment methods.

Several FR quality metrics take into account the lower-level aspects of the human visual system, such as contrast sensitivity, luminance masking, and texture masking [12, 13, 14, 15, 16, 17, 18]. These HVS-based quality metrics are allegedly more reliable than purely pixel-based FR-IQA methods, such as the peak signal-to-noise ratio (PSNR) and mean squared error (MSE). Other FR-IQA methods incorporate characteristics of the human visual system using feature extraction approaches [19, 20, 21, 22]. Webster *et al.* [23] proposed one of the first RR-IQA methods. Their method uses spatial and temporal features to assess the quality of videos.

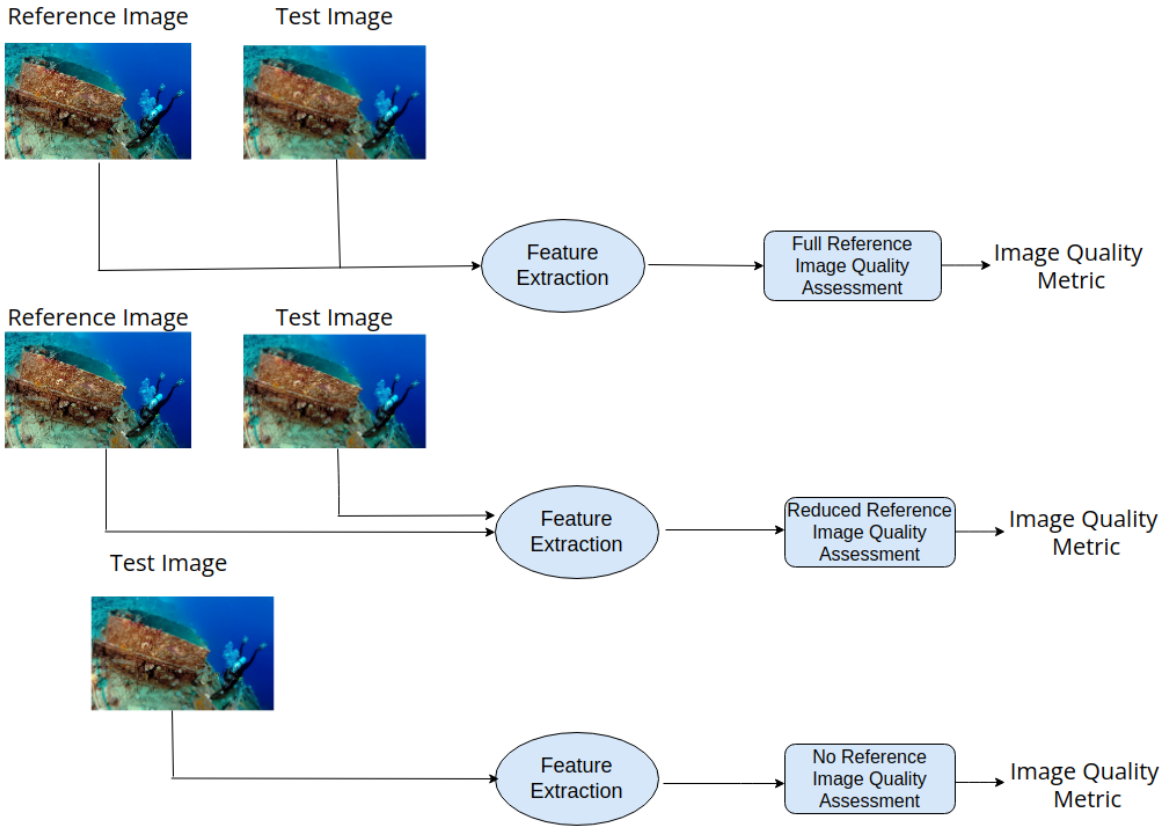

Figure 1.3: Types of image quality assessment.

For some multimedia applications, it is difficult to acquire information from visual reference content. In this scenario, NR quality metrics are the only available option. Unfortunately, these methods are generally less accurate than the FR methods, but they are often less complex. Some NR quality metrics use a distortion-specific approach [24, 25, 26, 27]. Despite the fact that NR-IQA methods have gained a lot of attention, their design remains a challenge [28, 29]. To further improve the reliability of IQA methods, a current research trend consists of investigating the impact of integrating visual attention into their design [30, 31, 32, 33]. This approach assumes that if a distortion occurs in an area that attracts the viewer's attention, it is more annoying than if it occurs in any other area. The algorithm weighs local distortions with local saliency.

Image enhancement algorithms have the goal of improving the image quality and, therefore, the usefulness of an image for a given task. Examples of image enhancement algorithms include gray contrast adjustments, sharpness enhancement (or deblurring), denoising, and color enhancement, which can be performed in the spatial or frequency domains. Figure 1.4 shows the basic work of the image enhancement technique to improve the quality of images according to the application scenario. However, although there are several image enhancement algorithms, there is no consensus on how to estimate the performance of these enhancement algorithms [6]. Since the final consumers of the resulting enhanced visual content are human viewers, the performance of these algorithms should take into account the perceived visual quality of the resulting enhanced

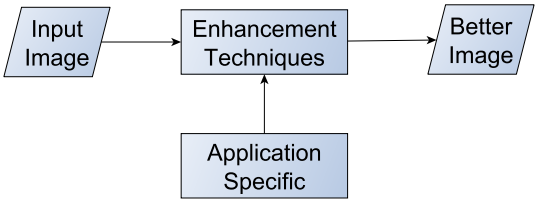

Figure 1.4: Image enhancement techniques.

images. We believe that a specific quality metric designed to assess the quality differences in enhanced or restored images can be used to estimate the success of an enhancement algorithm.

Unfortunately, although in recent decades much progress has been made in the area of image quality assessment, the design of metrics to estimate the quality of enhanced and restored images remains a challenge [34]. This is particularly true for underwater image application, where images often need to be restored due to the severity of degradations introduced by the underwater environment, which are often stronger than degradations in a regular (air) environment [35]. Therefore, there is a great need for quality metrics that can estimate the quality of enhanced and restored images [35]. In this thesis, our goal is to design metrics for this scenario. First, we have designed a quality metric based on texture operators and saliency. Second, we also designed a quality metric based on a deep learning convolutional neural network (CNN) architecture. We have developed a large and diverse image quality database for enhanced images. The database contains 35 SRCs images and 12 HRCs (different enhancement algorithms), resulting in a total of 455 test images. A crowdsourcing (online) psychophysical experiment was performed to obtain quality scores for these images. We also conducted a controlled laboratory experiment (onsite) and compared its results with the crowdsourcing (online) experiment. This database can be used to train and test image quality metrics, which are sensitive to enhancement algorithms that may increase or decrease perceived quality. We also generated the underwater image quality assessment database, which is helpful in designing the objective image quality assessment for underwater images.

1.2 PROBLEM STATEMENT

Image enhancement algorithms have the objective of improving image quality and therefore the usefulness of an image for a given task [36]. Many techniques have been proposed for image enhancement [37] these methods typically modify gray-level histograms, while other methods use local contrast transformations, edge analysis, or global entropy transformations, in all of these methods, there is no consensus on how to estimate the performance of these enhancement algorithms [36]. Exploring the underwater world has become increasingly important in recent years. Clear images in marine environments play an important role in underwater exploration and investigation, such as marine biodiversity monitoring, underwater rescue, underwater pipeline leak detection, underwater computer vision applications, etc. Poor visibility as a result of light

absorption and scattering poses a challenge to underwater image processing. Light absorption attenuates light energy, whereas scattering changes the direction of propagation of light. These effects drastically degrade the quality of underwater images, causing haze, loss of contrast, and color deviation. Therefore, enhancing underwater images is desirable and useful.

Underwater imaging applications require clear images achieved through enhancement and restoration techniques. Image restoration processes require physical degradation models that depend on parameters such as turbidity, time, and attenuation coefficient [35]. Image-enhancement methods are faster and do not require the calculation of parameters. The field of image enhancement research needs adaptive approaches and faster image enhancement techniques. This is particularly true for underwater image applications, where images often need to be restored due to the severity of the degradations introduced by the underwater environment. Therefore, there is a great need for quality metrics that can estimate the quality of enhanced and restored images. In this thesis, our goal is to design metrics for this scenario.

1.3 STATE OF THE ART OF UNDERWATER IMAGE QUALITY ASSESSMENT

The quality of underwater images is important for several applications, such as exploration of marine life, geological exploration, and archaeology. Capturing underwater images is a challenging process, due to the physical properties of underwater environments, which causes light to be attenuated when it transverses the water medium, either by absorption, scattering, or both. Absorption generates an energy reduction, while scattering changes the direction of light. Therefore, images captured underwater may have different types of degradation, including limited-range visibility, nonuniform lighting, low contrast, blurring, diminished color, bright artifacts, and noise. In other words, the visual aspect of underwater images may vary greatly depending on the characteristics of the water, including the types of particles present in the water and the depth of the water [38].

Currently, there are several methods in the literature to restore and improve the quality of underwater images [39, 40, 35]. Both image enhancement and image restoration algorithms have the goal of improving the overall quality of the acquired image. Image enhancement methods often use heuristic procedures to manipulate the image to achieve a pleasing visual aspect. Among the enhancement methods that have been used to improve the visibility of underwater images are histogram equalization algorithms [41, 42], retinex-based algorithms [43], particle swarm optimization (PSO) [44], and fusion-based algorithms [45].

On the other hand, to obtain the desired result, image restoration methods formulate a signal criterion that uses prior knowledge of the application/scenario and a degradation model to reconstruct or recover an undistorted image. There are several underwater image formation models. McGlamery and Jaffe proposed one of the first image formation models [46, 47]. This model has been used in the development of several underwater image acquisition systems. In 2006,

Trucco, Olmos, and Antillon [48] proposed a simplified version of the Jaffe-McGlamery model based on a self-tunning filter. To address the problem of blurred distortions in underwater images, which generally originates from light scattering in ripples and suspended particles, Hou *et al.* [40] incorporate optical properties to estimate lighting scattering parameters. Recently, several machine learning-based restoration algorithms [49, 50] have been proposed to dehaze and denoise underwater images.

The use of underwater images in computer vision and image processing applications often depends on the success of restoration and enhancement algorithms [51, 52, 53]. To determine the performance of these algorithms, we must estimate the quality of the restored/enhanced images as perceived by human viewers. In most cases, the success of underwater restoration and enhancement methods is measured using subjective qualitative observations and a few quantitative signal error measures [54, 55]. Most methods used to estimate the performance of these algorithms do not consider human perception or IQA methods. We believe that IQA methods are a viable option to estimate the quality of restored or enhanced underwater images.

For underwater scenarios, where a reference or undistorted image is not available, we must use NR-IQA methods to estimate the perceptual quality of restored and degraded images and determine if they are adequate for the target underwater engineering application. Although in recent decades great progress has been made in the area of image quality assessment, designing metrics to estimate the quality of enhanced and restored images remains a challenge [34]. As mentioned above, the final quality of underwater images depends on the type of underwater environment, which can introduce specific chroma, saturation, and contrast degradations [56, 57]. Since IQA methods are generally tuned to compression and transmission degradations, it is important to design methods that target the specific degradations of the underwater scenario. So far, few blind (no reference) image quality metrics have been proposed with the goal of evaluating the quality of underwater images [58, 1]. For example, Sanchez *et al.* [1] have proposed a underwater restoration algorithm that uses an NR-IQA method as a performance metric for the optimization algorithm.

Naturally, in the last decade, the use of machine learning and CNN techniques to design IQA methods has become extremely popular. By using deep learning CNN architectures, IQA methods are able to extract complex image features and obtain non-linear quality mappings, while requiring minimal knowledge of target domain. To the best of our knowledge, currently, there is no NR underwater IQA method based on CNN. In this paper, we propose an NR-IQA method for underwater images that is based on a deep learning CNN architecture. Our method uses a CNN architecture that is able to learn the features of the image more effectively and then estimate the image quality with greater accuracy. The proposed method employs small image patches as local quality measures for image denoising and reconstruction.

1.4 PROPOSED APPROACH

As mentioned earlier, in this work, our goal is to design IQA methods for enhanced and restored images. To better understand how the quality of enhanced images is affected by the algorithms parameters, the presence of degradations, and the type of content, we performed a psychophysical study in which subjects rated the quality of enhanced images with several types of enhancement algorithms, including color, sharpness, histogram, and contrast enhancements. The experiment was carried out online using a crowd-sourcing interface. This type of experiment has the advantage of making it possible to collect data from a large number of participants [59]. We also performed an experiment in a laboratory environment to serve as a control. Data from both types of experiments were compared. It is worth noting that there are very few databases of enhanced images in the literature.

In this work, we also studied the quality of images acquired in underwater scenarios. As mentioned earlier, this type of image may contain severe distortions due to light absorption and scattering, color distortion, poor visibility, and contrast reduction. Due to these degradations, researchers have proposed several algorithms to restore or enhance underwater images [60]. Naturally, one way to assess these algorithms' performance is to measure the quality of the restored/enhanced underwater images. Unfortunately, since reference (pristine) images are often not available, designing no-reference (blind) image quality metrics for this type of scenario is still a challenge. We propose two NR enhanced underwater image quality assessment (NR-UWIQA) methods.

The first method uses a multi-scale salient local binary pattern operator to estimate the quality of these images [61]. More precisely, we develop a method that combines information from salient maps and multiscale local binary pattern (MSLB) maps [62]. The proposed metric, named MSLB-UWIQA, was tested on the UID-LEIA database, showing good performance compared to other state-of-the-art methods. This approach will allow for a rapid and efficient evaluation of the results of restoration techniques, opening a new perspective in the area of underwater image restoration and assessment. The second method uses a deep learning architecture to estimate quality. We presented a light-weight NR-UWIQA method based on a CNN architecture. We trained the CNN model using patches of underwater images. Instead of feeding the model with patches selected randomly from underwater images, we chose patches that were most perceptually relevant by incorporating the properties of HVS name as edge and visual saliency. These HVS properties helped to calculate the regions of interest that were used as training data. The proposed method, named CNN-SP-UWIQA, has greater efficiency and robustness. Experimental results on the underwater image database demonstrate that our approach outperforms the state-of-the-art methods compared. We have developed an underwater image quality assessment database that helps us to design an underwater image quality assessment metric.

1.5 CONTRIBUTIONS

In this work, our achievements have been the following:

- We have developed an image enhancement database with images produced using several enhancement algorithms. We conducted a psychophysical experiment to collect subjective quality scores. Since there are few quality enhancement databases available in the literature, this work represents a contribution to the area of image quality.
- We have developed an NR-UWIQA metric based on the MSLB operator, which is simple, fast, and achieves good results.
- We have developed an NR-UWIQA based on a convolutional neural network. This metric performs well, outperforming state-of-the-art methods.
- We have developed a database for assessing the quality of underwater images that is diversified in both its contents and its restoration procedures.
- We have designed a generic IQA method that is capable of estimating the quality of restored and enhanced images in various scenarios.

1.6 ORGANIZATION OF THE DOCUMENT

This document is divided into six chapters. Chapter 2 describes the basic concepts that have been employed in this work with a brief literature review. In Chapter 3 we presents the a psychophysical study in which we analyze the perceptual quality of images enhanced with several types of enhancement algorithms, including color, sharpness, histogram, and contrast enhancement using a crowd sourcing framework. In Chapter 4 we presents the no reference underwater image quality assessment method that uses texture measurements to estimate underwater image quality. We also adopt a no reference underwater image quality metric with deep convolutional neural network. In Chapter 5 we develop a real underwater image quality assessment database which is helpful to design a no reference underwater image quality assessment metric which is useful to measure the quality of underwater images. Finally, in Chapter 6 we summarize the contributions of this work.

2 BASIC CONCEPTS AND LITERATURE REVIEW

In this chapter, we present a background to the basic concepts that have been used in the development of this work, including subjective and objective image quality assessment methodologies, texture operators, saliency models, underwater image processing, and the basics of deep learning (DL) architectures.

2.1 IMAGE QUALITY ASSESSMENT METHODOLOGIES

As mentioned earlier, IQA methods can be classified into two categories: objective and subjective methods [63]. Subjective quality assessment methods are psychophysical experiments in which participants rate the quality (or other attribute) of the images. Subjective experiments are the most precise way to estimate quality and are therefore considered a standard for the area. Unfortunately, these methods are time-consuming, expensive, and difficult to incorporate into multimedia applications. Objective image quality assessment methods are computer-based methods that can automatically predict perceived image quality. There is a great interest in this area given the difficulties of estimating the performance of multimedia applications. In this chapter, we introduce the basic concepts of these two types of IQA methods.

2.1.1 Subjective Quality Assessment Methodologies

In a psychophysical (subjective) experiment, a group of human subjects (participants) is asked to evaluate the quality (or other attribute) of a set of visual stimuli. Among the important aspects to consider when conducting a subjective experiment, some of the most important ones are the pool of participants, the setup (physical) conditions, the methodology used to collect the data and the statistical methods used to analyze the data. With respect to experimental methodology, there are several international recommendations that provide details on how to conduct quality experiments and obtain reliable results [64, 65, 66]. These recommendations suggest, among other things, standard viewing conditions, criteria for the selection of observers and test materials, the assessment procedure, and the data analysis methods.

It is important to include a pool of subjects with a diverse distribution of ages, gender, sexual orientation, geographic location, and cultural background [67]. For physical setup in subjective experiments, an important parameter is the viewing condition that affects the observers ability to perceive the attributes of the visual stimuli [68]. Most subjective experiments are performed in a laboratory environment, which is a strictly controlled environment designed to avoid errors and ensure reproducibility. Recently, researchers have been conducting subjective experiments online using crowdsourcing [69, 67, 70, 3]. In these experiments, researchers use commercial platforms

(Amazon Mechanical Turk and Microworkers) or social networks (Facebook, Twitter, LinkedIn) to recruit subjects to participate remotely in subjective studies. Crowdsourcing experiments have many advantages, including the ability to recruit large numbers of participants. Although crowdsourcing experiments are very practical, collected scores are considered less reliable due to the challenges and limitations of these experiments [71].

In terms of rating scales, the methodologies can be divided into direct and indirect rating methods. Direct scaling methods collect the opinions of observers regarding each distinct stimulus on the ratio scale. The scale value depends on the selected procedure. Once all the scores given by all observers have been collected, the processing (including outlier detection and averaging) is applied directly to the raw data [72]. The advantage of direct rating is that the data from individual observers are taken directly from the scale, which simplifies the processing and interpretation steps. Another advantage of the direct scaling procedure is that the scores of individual observers can be placed directly on a clearly defined rating scale, which simplifies the subsequent processing and interpretation. However, indirect procedures typically provide higher discriminatory power and can be less complicated and tiring for observers. Several works have shown that indirect scaling methods need a smaller number of subjects to provide the same reliability as direct scaling procedures [72].

Generally, there are three types of experimental methodologies to conduct subjective experiments: single-stimulus methods, double-stimulus methods, and stimulus comparison methods [73]. These methods generally have different applications, and the choice of a specific method depends on the context, the purpose, and where the test will be performed in the data development process. In single stimulus (SS) methods, test images are randomly displayed for a fixed amount of time. After that, their disappearance from the screen, the observer is asked to score their level of quality on a scale of 1 to 5. This quality scale is known as the Absolute Category Rating Scale (ACR) [73], which is shown in Figure 2.1(a). A common variation of the SS methodology is the Single Stimulus Absolute Category Rating Scale Hidden Reference (ACR-HRR), in which the (unprocessed) reference sequence is included in the experimental sessions without any identification. Another variation is the Single Stimulus Continuous Quality Evaluation (SSCQE), in which the evaluation is performed continuously. For each test stimulus, the SS experiments output a mean opinion score (MOS) value, which is computed by taking the average of the ratings for all observers:

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

where S_i is the score of an observation and N is the total number of observations. The MOS can also be used to calculate the Difference Mean Opinion Score (DMOS), where the MOS for the reference is then subtracted from the MOS for the other images. This generates a DMOS for each image. The DMOS value for a particular test represents the subjective quality of the test relative to the reference.

ITU-R Rec.BT500-13 details how to process the data gathered from subjective experiments.

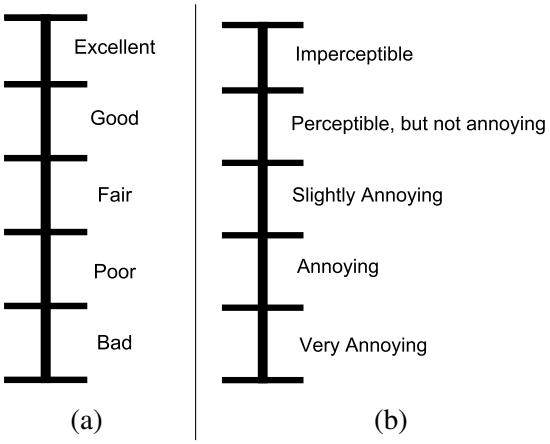

Figure 2.1: (a) Quality Category Rating Scale, (b) Impairment Category Rating Scale.

It is common to compute confidence intervals (CI) of MOS values to show the variability of subjective quality scores. The confidence level refers to the certainty that the confidence interval contains the true population parameter when drawing numerous samples. The most common confidence limits are 95% and 99%. This means that 95% of random samples drawn with a 95% confidence interval will contain the true parameter. The confidence interval is computed using the following equation:

$$\text{CI} = \bar{x} \pm z\frac{s}{\sqrt{n}}, \tag{2.2}$$

where $\bar{x}$ is the sample mean, s is the sample standard deviation, n is the sample size, and z is the number of standard deviations from the mean.

In Double Stimulus Methods (DS), test and reference sequences are presented simultaneously to the participants, who evaluate their quality by comparing them. Similarly to the SS method, the evaluation can be performed for short or continuous sequences, but as a consequence the length of the experimental session is larger. Variations in DS methodologies include the Double Stimulus Impairment Scale (DSIS), the Double Stimulus Continuous Quality Scale (DSCQS), and the Simultaneous Double Stimulus for Continuous Evaluation (SDSCE). In DSIS, also called Degradation Category Rating (DCR), the observer is first shown an unimpaired image, and then the same content is degraded. The observer is asked to rate the distortion in the second image in relation to the first image. The method uses a five-point impairment scale, shown in Figure 2.1(b). In the DSCQS method, the observer rates the quality of the pair of test image (probably impaired) and reference image using a continuous scale ranging from 0 to 100. However, the order of presentation of the reference image is randomly changed throughout the experimental sessions. The final result of a DS experiment is given in terms of differential mean opinion scores (DMOS) values. The DMOS is calculated taking the difference between the MOS corresponding to the reference image and the MOS corresponding to the test image [64].

In terms of stimulus comparison methods, although there are many of these methods, we mention here the ordering by force-choice pairwise comparison and the pairwise similarity judgment [64]. In the first method, two images of the same content are displayed, and observers are

asked to choose the image with the highest quality. Observers are always required to choose one image, even if they perceive them as having the same quality. In the second method, observers are shown the pair of images. But in this case, they are asked not only to choose the image with higher quality, but also to indicate the level of difference between them on a continuous scale.

The traditional subjective experiments mentioned above require the following:

1. a diversified pool of participants;
2. a laboratory environment, with physical conditions that adhere to the recommendations (e.g., Rec. ITU-R BT.500 [74] and Rec. ITU-T P.910. [73]);
3. a dataset of test stimuli that the participants will evaluate.

In recent years, crowd-sourcing experiments have been used in various application areas as a cost-effective substitute for subjective experiments in a laboratory setting [59, 75, 76, 77, 78, 79]. In other words, instead of conducting the subjective experiment in a controlled laboratory setting, researchers use dedicated platforms (e.g., Amazon Mechanical Turk[1] and Microworkers[2]), social networks (Twitter, Facebook, Linkedin) or email campaigns to recruit subjects to participate remotely in subjective studies. Crowdsourcing experiments have many advantages, including the ability to recruit large numbers of participants. Additionally, crowdsourcing campaigns typically require a short implementation time and can collect data from a variety of subjects with different backgrounds, testing environments, and devices.

2.1.2 Objective Quality Assessment Methods

As mentioned earlier, subjective experiments have several disadvantages, including the fact that they are time-consuming, expensive, and require access to a panel of naive and diverse pool of human observers. Objective quality methods, on the other hand, are algorithms (implemented in hardware or software) that automatically estimate the quality of an image [80, 81]. These methods are designed and tested using subjective quality scores given to the visual content as ground-truth. Recently, the area of image and video quality has made great progress, with the performance of visual quality metrics improving considerably [6, 82, 83]. For this reason, objective quality metrics, which are basically algorithms that are able to estimate quality automatically, are very important in real-time multimedia applications.

Over the years, numerous objective quality metrics have been proposed. Based on the degree of availability of the original visual content, quality metrics are categorized into FR metrics, which need the reference content to perform the quality assessment, RR metrics, which require only certain attributes (information about edges, histogram, etc.) of the reference content, and NR or "blind" quality metrics, which assess image quality using only the test image without requiring any information about the reference content. However, there are still many challenges

[1] https://www.mturk.com
[2] https://microworkers.com

in this area, including the design of objective quality metrics to enhance visual content. Since most quality metrics are designed to capture visual distortions, they cannot quantify the quality changes produced by enhancement algorithms. Therefore, metrics that can automatically estimate the quality of enhanced images and videos are needed. One of the challenges in developing quality assessment methods for enhanced visual content is the lack of quality databases containing enhanced images and their respective subjective quality scores (ground truth).

Most FR metrics measure the "fidelity" or similarity of the test image compared to the reference image. The reference is assumed to be the version of the content of the best possible quality. FR image quality assessment (FR-IQA) methods provide the most reliable quality estimates [84]. One type of FR-IQA metrics is signal-based metrics or pixel-based metrics, which compare the original and processed images at the level of pixel intensities. These types of metrics are very popular because of their simplicity, low computational complexity, and clear mathematical meaning. However, these methods do not always correlate well with subjective quality scores, since human observers do not perceive quality as pixel differences [85].

An example of a signal-based IQA metric is the mean squared error (MSE), which is defined as:

$$\text{MSE}(I_R, I_P) = \frac{1}{X \cdot Y} \sum_{x=1}^{X} \sum_{y=1}^{Y} \left(I_{\text{R}}(x,y) - I_{\text{P}}(x,y) \right)^2, \tag{2.3}$$

where I_R is the reference image, I_P is its processed version (test), and X and Y represent the width and height of the image in pixels, respectively. MSE measures the error between this pair of images, with higher values of MSE corresponding to lower qualities and lower values corresponding to higher qualities. The most popular pixel-based metric is the peak signal-to-noise ratio (PSNR), which is computed as follows and expressed in decibels (dB):

$$\text{PSNR}(I_R, I_P) = 10 \log_{10} \frac{\left(2^B - 1\right)^2}{\text{MSE}(I_R, I_P)}, \tag{2.4}$$

where B stands for the number of bits used to express each pixel intensity value in the image, that is, $B = 8$ for 8-bit images.

Many IQA metrics have been proposed that take into account the low-level characteristics of HVS[86]. Among the most popular IQA metrics is the SSIM index proposed by Wang [83], which attempts to quantify the visible difference between a distorted image and a reference image. This index is based on the Universal Image Quality (UIQ) index [83]. The algorithm defines the structural information in an image as those attributes that represent the structure of the objects in the scene, independent of the average luminance and contrast. The index is based on a combination of luminance, contrast, and structure comparison. The comparisons are done for local windows in the image, and the overall IQ is the mean of all these local windows. The SSIM is defined as:

$$\text{SSIM}(x,y) = \frac{\left(2\mu_x\mu_y + C_1\right)\left(2\sigma_{xy} + C_2\right)}{\left(\mu_x^2 + \mu_y^2 + C_1\right)\left(\sigma_x^2 + \sigma_y^2 + C_2\right)}, \tag{2.5}$$

where μ_x and μ_y are the mean intensities of windows x and y, respectively, σ_x and σ_y are the standard deviation of windows x and y, respectively, and σ_{xy} is the cross standard deviation of these windows. C is a constant defined as

$$C_1 = (K_1 L)^2 , \tag{2.6}$$

where L is the dynamic range of the image and $K_1 \ll 1$. C_2 is similar to C_1 and is defined as:

$$C_2 = (K_2 L)^2 , \tag{2.7}$$

where $K_2 \ll 1$. These constants are used to stabilize the division of the denominator. SSIM is then computed for the entire image as:

$$\mathrm{MSSIM}(I_R, I_P) = \frac{1}{W} \sum_{j=1}^{W} \mathrm{SSIM}\left(x_j, y_j\right), \tag{2.8}$$

where I_R is the reference image, I_P is its processed version (test), x_j and y_j are the image contents of the j-th local window, and W indicates the total number of local windows.

Another FR-IQA metric is Visual Information Fidelity (VIF) [87], which uses three models to calculate the quality of distorted images. Both metrics are known as Natural Scenes Statistics (NSS)-based IQA metrics. Some NSS-based metrics model the subband filtered coefficients of the reference image. Each coefficient is expressed as a random variable employing a Gaussian-scale mixture (GSM). A very interesting feature of VIF is its ability to recognize if the test image is of superior quality and output a VIF index greater than one. Although this works mostly for images with increased contrast, it is the first step towards designing IQA methods for enhanced images.

The feature similarity index (FSIM) [63] is an IQA metric extracts and compares low-level features from reference and test images. FSIM has been extended to consider information about color by converting the original colorspace YIQ [63]. FSIM maps the features and measures the similarities between two images. To describe FSIM we need to describe two criteria more clearly: the phase congruency (PC) and the gradient magnitude (GM) [88]. PC is a method for detecting image features that is invariant to light variation and contrast. It is also capable of detecting more interesting features, stressing the features of the image in the frequency domain. We use the horizontal and vertical Sobel operators G_x and G_y, respectively[89], and compute the magnitude of the overall gradient of an image I:

$$G_x(I) = \begin{bmatrix} 1 & 0 & -1 \\ 2 & 0 & -2 \\ 1 & 0 & -2 \end{bmatrix} * I, \qquad G_y(I) = \begin{bmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{bmatrix} * I. \tag{2.9}$$

Next, obtain the overall gradient:

$$G_I = \sqrt{G_x^2 + G_y^2}. \tag{2.10}$$

Then, FSIM calculates the GM maps of two images, f_1 (test image) and f_2(reference image): G_1 and G_2, respectively. Next, it computes the PC maps, I_R and I_P, from these two images: PC_R and PC_P. Then, it computes the similarity of the PC maps:

$$S_{PC}(I_R, I_P) = \frac{2 \cdot PC_R PC_P + T_1}{PC_R^2 + PC_P^P + T_1}, \tag{2.11}$$

and of the GM maps:

$$S_G(I_R, I_P) = \frac{2G_{I_R} G_{I_P} + T_2}{G_{I_R}^2 + G_{I_P}^2 + T_2}. \tag{2.12}$$

In these equations, T_1 is a positive constant that increases the stability of S_{PC}. T_2 is a positive constant that depends on the dynamic range of the gradient magnitude values. The above equations describe the measurement to determine the similarity of two positive real numbers, outputting a number ranging from 0 to 1.

Finally, S_{PC} and S_G are combined together to calculate the similarity S_L of I_R and I_P:

$$S_L(I_R, I_P) = [S_{PC}(I_R, I_P)]^\alpha \cdot [S_G(I_R, I_P)]^\beta, \tag{2.13}$$

where the parameters α and β are used to adjust the relative importance of the PC and GM features. We set $\alpha = \beta = 1$ for convenience. The FSIM index between I_R and I_P is defined as follows.

$$\mathrm{FSIM}(I_R, I_P) = \frac{\sum \mathbf{x} \in \Omega S_L(\mathbf{x}) \cdot PC_m(\mathbf{x})}{\sum \mathbf{x} \in \Omega PC_m(\mathbf{x})} \tag{2.14}$$

where Ω means the whole spatial domain of images I_R and I_P.

The Most Apparent Distortion (MAD) [90] is based on the assumption that HVS judges the quality of the image according to the degree of distortion in the image. When the image is only slightly distorted, observers tend to search for distortions, which is called a "detection-based strategy." Then, the model combines the local masking model computed in the spatial domain with the local MSE calculated in the perceived luminance domain. The low-level properties of HVS (contrast sensitivity, nonlinear perception of luminance, luminance, and contrast masking) are combined into a map that contains the locations of visible distortions. Using this map, the visibility-weighted MSE map is computed and combined to form a single scalar value using the L_2 norm.

In the case of heavily distorted images, the HVS does not have to look for distortions and an "appearance-based strategy" is used. Here, the distortion is expressed as the extent to which the appearance of the image content is degraded, which is computed using the local statistics of multi-scale log-Gabor filter responses. In other words, the image is decomposed with log-Gabor filters with four orientations and five different scales. Block-based statistics, such as variation, skewness, and kurtosis, are obtained from every decomposition. These are then combined into the statistical difference map, which is then again collapsed into a single scalar number. The final

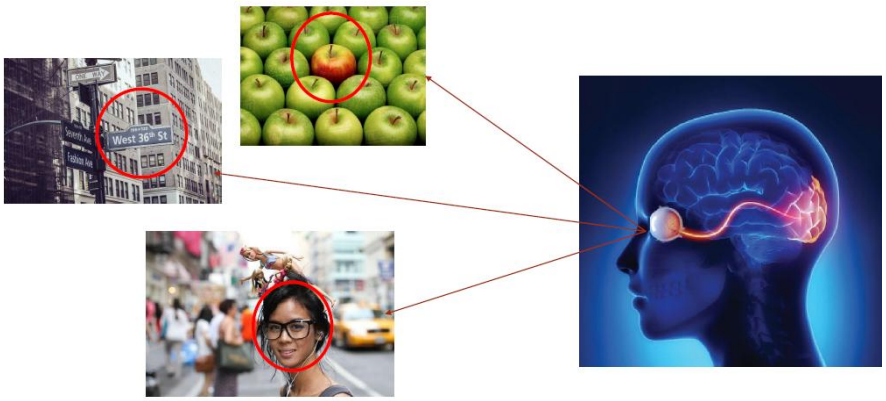

Figure 2.2: Example of salient regions (red circles on the left images) that attract the Human Visual System.

output by MAD is as follows:

$$\text{MAD} = (d_{\text{detect}})^{\lambda} (d_{\text{appear}})^{(1-\lambda)}, \tag{2.15}$$

where d_{detect} and d_{appear} denote the values obtained by the "detection-based strategy" and "appearance-based strategy", respectively, and λ is the weight chosen according to the overall level of distortion.

For heavily distorted images, λ should be close to 1. No optimal procedure for the selection of λ is defined, but the authors achieved good results with λ calculated as follows:

$$\lambda = \frac{1}{1 + \gamma_1 (d_{\text{detect}})^{\gamma_2}}. \tag{2.16}$$

MAD is used on the luminance component of the image when the calculation of d is computationally demanding.

2.1.3 Visual Attention for Image Quality Assessment

Visual attention (VA) is a mechanism of HVS. When observing a scene, the human eye filters the large amount of visual information available and focuses on selected (salient) regions [91]. This selection process is actively controlled through oculomotor mechanisms that allow the gaze of attention to hold on a particular location (fixation) or to shift to a preferred location when sufficient information has been collected from the current focus (saccades). Fixations are instinctively concentrated on highly informative areas; as a consequence, the amount of data to be further processed by the brain is minimized, but maximizing the quantity of useful information. Figure 2.2 represents the VA mechanism, with the red circles in the images depicting the salient regions captured with the image eye fixations captured with an eye tracker.

Two different mechanisms operate in VA: Bottom-up and Top-down. Bottom-up attention

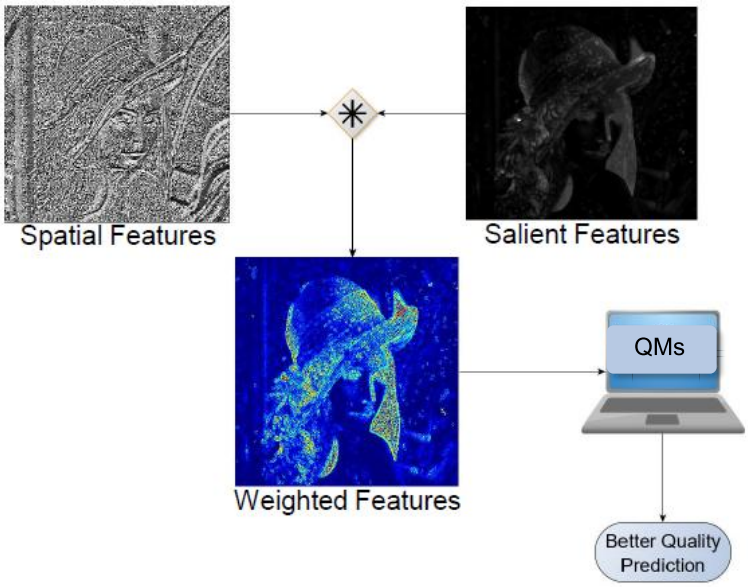

Figure 2.3: Saliency Weighted Quality Assessment.

is stimuli-driven and is based on intrinsic features of the input image such as orientation, color, intensity, and motion. Bottom-up attention is the result of simple feature extraction across the whole visual field. Therefore, a highly salient region of an input visual content can capture the focus of human attention. For example, flashing points of light on a dark night, sudden motion of objects in a static environment, and red followers on a green background can involuntarily and automatically attract human attention. Top-down attention is a task-driven mechanism and refers to the set of processes used to bias visual perception based on the task or intention. For example, an observer is assigned a task to find a black pen in a scene of a room crowded with many other things. Based on the prior knowledge, experience, and current objective of the observer, which are mostly controlled by the high-level cortex, the selection of a region is established. Bottom-up attention pops out only the candidate regions where targets are likely to appear, while top-down attention can depict the exact position of the target.

Algorithms that detect salient regions automatically, without human intervention, are called saliency models. For example, Itti, the graph-based visual saliency (GBVS), and the Boolean Saliency Map (BMS) models [92, 93] are the most common and widely used bottom-up saliency models. They use conventional programming strategies to compute saliency, in which saliency is generated in three steps. In the first step, *extraction*, lower-level image features (contrast, luminance, and textures) are extracted and organized into a vector format. In the second step, *activation*, saliency maps are generated from each feature vector. Finally, in the third step, *combination*, saliency maps are combined into one final saliency map. For example, the BMS first generates all possible Boolean maps of a picture frame. Then, it applies a threshold to them to create activation maps. Finally, a saliency map is generated by computing the mean of all activation maps.

Visual attention plays an important role in IQA. Any distortion that occurs in a salient area

is more important to the overall perceived quality. Therefore, the distortion that occurs in salient areas should be treated differently from the distortion in less salient areas. For this purpose, visual attention can be used to improve the accuracy of IQA metrics to make quality estimates more accurate. Recently, a variety of conventional image quality metrics [94, 95] and video quality metrics [18, 17, 32, 33, 96, 97, 98, 99], have incorporated the saliency information to improve their quality predictions. In most existing works, visual attention is used as a weighting factor to spatially pool objective quality estimates from a quality map [100]. A weighted salient-based IQA metric can be computed as follows:

$$\varepsilon_{sal} = \frac{\sum_{x,y} \varepsilon(x,y) \cdot \omega(x,y)}{\sum_{x,y} \omega(x,y)}, \tag{2.17}$$

where x and y are the horizontal and vertical spatial indices, $\varepsilon(x,y)$ is the error map computed using an FR-IQA, and $\omega(x,y)$ is the saliency map calculated using a salient model that outputs values between 0 and 1, with 1 corresponding to highly salient pixels and 0 to non-salient pixels. ε_{sal} denotes the salient weighted IQA measure, which estimates quality and gives a higher importance to errors in highly salient areas and a lower importance to errors in less salient areas. Figure 2.3 depicts this process of weighing salient regions. In this figure, the FR-IQA metric is the SSIM and the saliency map is the BMS.

The work of Zhang *et al.* [31] presented detailed statistical evaluations on the performance of saliency-weighted quality metrics for both images and videos. More recently, a CNN-based NR-IQA method [101] has been proposed that incorporates saliency information in image quality prediction.

2.1.4 Image Quality Performance Metrics

The effectiveness of objective quality metrics is generally quantified by the degree to which its quality prediction agrees with human judgements (MOS or DMOS), that is, by comparing the predicted quality scores with subjective quality scores. Statistical measures are often used as performance metrics. In visual quality assessment, the most common and widely used performance metrics are Spearman's rank order correlation coefficient (SROCC) and Pearson's linear correlation coefficient (LCC).

LCC measures the degree of relationship between linearly related variables and is calculated as follows [102, 103]:

$$r_{x,y} = \frac{\sum_{i=1}^{n}(x_i - x)(y_i - y)}{\sqrt{[\sum_{i=1}^{n}(x_i - \bar{x})^2]\,[\sum_{i=1}^{n}(y_i - \bar{y})^2]}}, \tag{2.18}$$

where $r_{x,y}$ LCC between x and y, n is the number of observations, x_i is the value of x at i^th observation and similarly y_i is the value of y at i^th observation. SROCC is a nonparametric measure that is used to measure the degree of association between two variables and is calculated

as follows [102, 103]:

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}, \tag{2.19}$$

where r_s denotes SROCC, d_i is the difference between the ranks of the corresponding variables and n is the number of observations. The difference between SROCC and LCC is that SROCC benchmarks the monotonic relationship between two variables, while LCC benchmarks the linear relationship. The PLCC and SROCC values closer to 1 represent the better performance of the quality metric.

2.2 UNDERWATER IMAGE PROCESSING

The underwater environment is the area immersed in water in a natural or artificial body of water such as an ocean, sea, reservoir, river, or aquifer. The underwater environment is considered the origin of life on Earth and is the most important environmental region for sustaining life and the natural habitat of most living organisms. Many human activities are carried out in accessible areas of the underwater world. Therefore, it is important to understand the characteristics of the underwater imaging model to conduct research in different areas.

The secrets of deep-sea ecosystems can be unlocked to provide new sources for the development of pharmaceuticals, food and energy resources, and renewable energy products. Underwater imaging research has increased significantly over the past decade. This is mainly due to human dependence on valuable underwater resources. Effective exploration of the underwater world is only possible if there are excellent methods to enhance underwater images. Underwater imagery analysis is important for marine resource exploration, marine ecology research, deep-sea asset monitoring, and naval military applications [104, 105]. The use of available underwater resources is of great importance to humans. For this reason, remotely operated underwater vehicles (ROVs) and autonomous underwater vehicles (AUVs) equipped with a high-quality imaging system are needed to effectively explore the underwater environment. The low quality of underwater images leads to the failure of the computer systems used to visual inspection the images. Therefore, the development of underwater enhancement techniques to perform challenging underwater imaging tasks is extremely important. In recent decades, much research has been conducted in the field of underwater image recovery and enhancement.

Underwater image processing deals with the refinement of image information necessary for human and machine interpretation and processing [106]. Underwater research in the areas of detection of cracks in underwater pipelines, exploration of flora and fauna, and marine archaeology is largely based on clear and unambiguous underwater images. Complete investigation of underwater applications is dependent on the quality of the captured underwater images. In general, the quality of underwater photographs depends on numerous aspects, such as limited visibility, uneven illumination, unwanted signals such as noise, and degrading colors. Underwater image enhancement can be broadly divided into two categories, namely image restoration and

image enhancement. The underwater image enhancement does not use a physical model for the enhancement task. It uses qualitative and subjective criteria to produce a visually pleasing image. Enhancement-based approaches are simpler and faster than model-based approaches [107].

2.2.1 Challenges of Underwater Imaging

Underwater imaging systems consist mainly of optical cameras or specialized equipment to capture underwater images [108]. Except for optical cameras, all other methods have their own limitations, such as restricted field of view, depth limitations, complex operations, etc. ROV control is complex due to unfamiliar nonlinear hydrodynamic effects, lack of an accurate model of ROV dynamics, and uncertainty of parameters. The underwater research requires highly skilled divers, it is an expensive proposition. Each investigation may require divers and supervisors on call for a single mission. Additionally, only a limited amount of time can be spent in an underwater medium, especially when a diver conducts an investigation. This results in an increase in the time required for exploration. This disadvantage can be overcome largely by using underwater image enhancement techniques.

When light propagates through water, the optical properties of water adversely affect underwater imaging. The refractive effect in an underwater medium makes it much more difficult to focus on the object being photographed, resulting in a blurred image. Light and color absorption significantly affects the visual quality of underwater images. To overcome this limitation, it is important to conduct research to develop efficient and powerful enhancement methods. The medium of water is a natural light source that absorbs a significant amount of light that passes through it. For every 10 m depth under water, half the light is lost. This means that at a depth of 10 m, we have only 50% of the light we had on the surface and at a depth of 20 m, only 25% [106].The availability of light is not always the same; it depends on the time of day. The surface of water reflects the least amount of light when the Sun is directly overhead. Weather plays a crucial role in the availability of light. In stormy weather, turbulent water has a significant effect on light conditions. Color absorption is another major challenge, as a result of the absorption of wavelengths of light. Blue and green colors have longer wavelengths compared to red colors and therefore extend further into the depths, which is why underwater images have blue and green hues.

Absorption and scattering result in blurring, reduced contrast, and a general loss of image quality. This problem is exacerbated when there is high turbidity underwater or when strong artificial light sources are used. Artificial light sources cause uneven illumination of the scene, resulting in reflections that obscure image details and create bright spots [35]. This can lead to misinterpretation of the data. The fluorescence of biological objects and the presence of macroscopic particles in the water cause the degradation of underwater images. Underwater imaging provides measures to overcome some of these challenges by making visual information part of the quantitative assessment. The application of efficient image-based methods will provide accurate quantitative information with a minimum of human oversight to complement visual inspection techniques and improve reliability.

Images acquired in underwater scenarios may contain severe distortions due to light absorption and scattering, color distortion, poor visibility, and contrast reduction. Due to these degradations, researchers have proposed several algorithms to restore or enhance underwater images. One way to assess these algorithms' performance is to measure the quality of the restored/enhanced underwater images. Unfortunately, since reference (pristine) images are often not available, designing no-reference (blind) image quality metrics for this type of scenario is still a challenge. In fact, although the area of image quality has evolved a lot in the last few decades, estimating the quality of enhanced and restored images is still an open problem. In this work, we present two no-reference image quality evaluation metrics for enhanced underwater images (NR-UWIQA). The first metric uses an adapted version of the multiscale salient local binary pattern operator to extract image features, while the second metric uses a deep learning architecture. Both proposed metrics were tested on the UID-LEIA database, presenting good accuracy performance compared to other state-of-the-art methods.

2.2.2 Underwater Imaging Model

The presence of dust particles leads to the phenomenon of scattering in the underwater medium. The light reflected from the outside of the object reaches the camera, and a scattering effect occurs when the light interacts with suspended particles present in the imaging medium. The two types of scattering that affect underwater images are forward scattering and backward scattering [104]. Forward scattering E_f occurs when the light reflected from the object is scattered as it travels in its direction before reaching the camera, which in turn results in a blurred image. Backward scattering occurs when the reflected light reaches the camera directly before it reaches the scene to be illuminated. This results in low contrast and a haze-like effect on the image [105, 107].

According to Lambert-Beer's empirical law, the decrease in light intensity depends on the properties of the medium through which the light travels [35]. The intensity decreases exponentially as it spreads over water. This loss of intensity is called attenuation. It results from the effects of absorption, which causes a loss of light energy, and scattering, which causes a change in the direction of the electromagnetic energy [52]. The phenomenon of absorption and scattering leads to attenuation of light. Light attenuation is important in underwater images because it results in a hazy effect that hinders imaging applications in the marine environment. Light attenuation limits visibility to about 20 m in clear water and 5 m in turbid water. The absorption of light in water varies with wavelength. The different colors of light disappear with increasing water depth. Red light is absorbed first because of its shortest wavelength. The blue color penetrates the farthest into the water medium due to its longest wavelength, hence the bluish coloration in underwater images [106, 35].

Underwater imaging models are often used for image restoration algorithms. The Jaffe-McGlamery underwater imaging model is a well-known imaging model proposed by Jaffe and McGlamery. Figure 2.4 illustrates the fundamentals of this model, where light captured in the underwater medium is represented as the linear superposition of a direct component, a forward

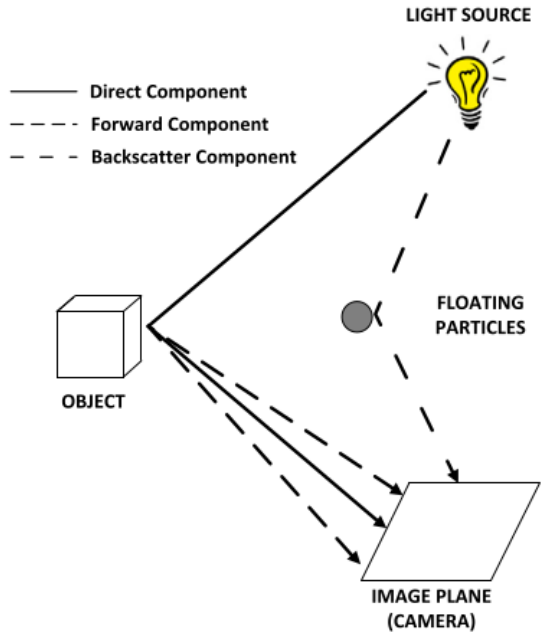

Figure 2.4: Illustration of the Jaffe-McGlamery model for underwater light propagation [1].

scatter component, and a backward scatter component. The model is based on the facts of linear superposition and modeling of medium water. The total irradiance E_t entering the camera consists of a linear combination of these three components, given by:

$$E_T = E_d + E_f + E_b, \tag{2.20}$$

where E_d is the direct component, E_f is the forward scattered component, and E_b is the backward scattered component.

2.2.3 Underwater Image Enhancement

The main goal of image enhancement is to process an image so that the resulting image is of better quality than the original image. Image enhancement techniques can be used to improve the interpretability of images for humans or computer algorithms. To achieve these goals, image enhancement algorithms adjust the attributes of the input image according to specific tasks. There are many digital image enhancement techniques, and the appropriate choice depends on the imaging application, the task at hand, and the viewing conditions. Image enhancement methods can be divided into the following two categories: spatial domain methods and frequency domain methods. In spatial domain techniques, to achieve the desired enhancement, the processing is performed directly on the image pixels. In frequency-domain methods, the image is first converted to the frequency domain, where processing is performed.

Image enhancement can be used in all applications in which images are consumed, such as multimedia transmission applications, medical imaging, satellite image analysis, and underwater imaging. Image enhancement algorithms simply transform an image I into an image I_h using

a transform T. Representing the pixels in the images I and I_h by r and s, respectively, this transformation can be expressed by the following expression:

$$s = T(r),$$

where T is a transformation that maps the pixel value r into a pixel value s.

As mentioned earlier, the design of IQA metrics is an important step for measuring the performance of enhancement techniques. Moreover, observer-specific factors, such as HVS properties and the observer's experience, will introduce a great deal of subjectivity into the choice of image enhancement methods [109]. In the particular case of underwater image processing has received considerable attention within the last decades, showing important achievements. In this work we review some of the most recent methods that have been specifically developed for the underwater environment. This is primarily due to the dependence of human beings on the valuable resources that exist in underwater. Effective work of exploring the underwater environment is achievable by having excellent methods for underwater image enhancement. These techniques are capable of extending the range of underwater imaging, improving image contrast and resolution. After considering the basic physics of the light propagation in the water medium, we focus on the different algorithms available in the literature. The conditions for which each of them have been originally developed are highlighted as well as the quality assessment methods used to evaluate their performance. Fig. 2.5 illustrates an example of pristine underwater images with their enhanced versions.

2.2.4 Database for Underwater Images Quality Assessment

However, currently available IQA methods have shown little success when blindly evaluating the quality of underwater images. In this work, we present no-reference (underwater) image quality assessment (NR-UWIQA) methods, which will be described in detail in the next chapters. To design these metrics, we must use real-world underwater image datasets. An example of an underwater dataset is the Fish4Knowledge dataset [110], which is used for the detection and recognition of targets. The SUN dataset [111] is used for object detection, the MARIS dataset [112] is used for autonomous marine robotics, and the SEA-thru [113] dataset is used for range maps. However, existing datasets do not provide ground-truth or reference images, which makes it difficult to design image quality metrics for these types of applications. Recently, Sanchez *et al.* proposed UID-LEIA (Underwater Image Database of the Laboratory of Embedded Systems and Integrated Circuit Applications) [1]. Figure 2.6 shows a sample of the 45 reference images and 135 distorted underwater images contained in this dataset. In addition to the images, UID-LEIA also contains subjective quality scores for all of these images, making it possible to use this dataset in the design of underwater IQA methods. We have also developed a real-world underwater image dataset, which is helpful in designing a no-reference metric.

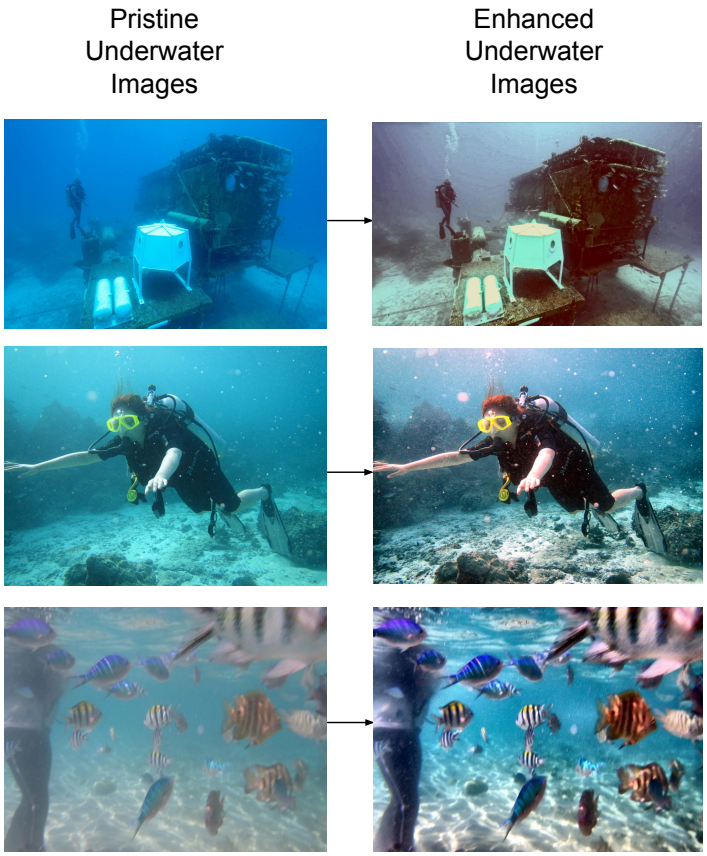

Figure 2.5: Illustration of pristine underwater images with their enhanced versions.

2.3 BRIEF INTRODUCTION TO MACHINE LEARNING

Machine Learning (ML) is a subset of the knowledge domain of Artificial Intelligence (AI). ML is defined as "algorithms that process data, learn from those data, and then apply what they have learned to make decisions" [114]. Alpaydin [115] mentions that machine learning can be used to program computers using optimization performance criteria and example data. We have a model defined up to some parameters, and learning is the execution of a computer program to optimize the parameters of the model using the training data or past experience. The model may be *predictive*, to make predictions, or *descriptive*, to gain knowledge from the data, or both [115]. For example, by analyzing sample face images, a learning program captures the pattern specific to each person and then recognizes them by checking for this pattern in a given image. This is an example of a pattern recognition problem [115]. Based on the types of applications, ML algorithms can be divided into four categories: supervised, unsupervised, semi-supervised, and reinforcement learning. As shown in Figure 2.7, supervised learning is a predictive model that processes labeled data to achieve a specific task.

For example, regression and classification are supervised learning applications. Some of the most popular regression methods are as follows: Random Forest Regressor (RFR), Support Vector Machine (SVM), Linear Regression (LrR) and Logistic Regression (LgR). Unsupervised learning

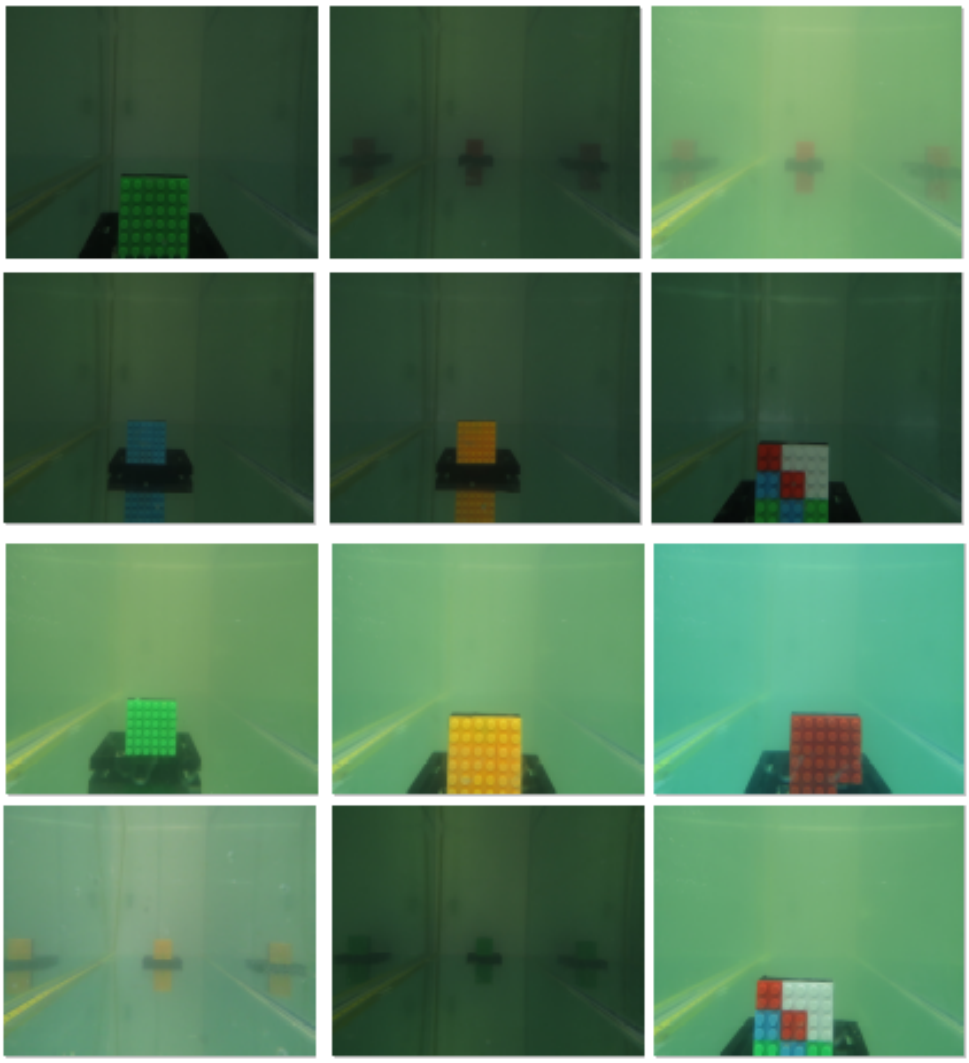

Figure 2.6: Sample underwater images from UID-LEIA database.

is a descriptive model that processes unlabeled data. The model learns from the hidden structure of the data. An application of unsupervised learning methods is to learn associations of different attributes of data. Semi-supervised learning is the combination of supervised and unsupervised learning methods. For example, learning from unstructured data to define tags and types of content in text classification problem is one example of an application of semi-supervised methods. In reinforcement learning, the model assesses the policies (rules) and learns from past good action sequences to be able to generate a policy. For example, in some applications, the output of the system is a sequence of actions. In such a case, a single action is not important; what is important is the policy, which is the sequence of correct actions to achieve the goal. An action is good if it is part of a good policy. Robotic cars are one of the reinforcement learning methods.

In IQA metrics, the most commonly used ML methods are SVR, LrR, and RFR. SVR is a kernel-based regression method that uses variants of kernel functions for learning. For example, CORNIA [116], CQA [117], SSEQ [118], BRISQUE [119], LTP [120], DIIVINE [121], MLBP

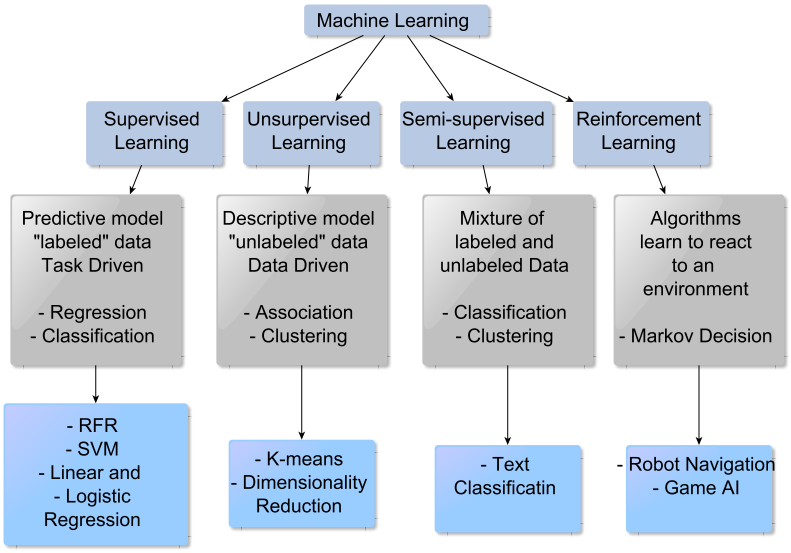

Figure 2.7: Categories of machine learning algorithms

[2], and GWH-GLBP [122] are FR-IQA metrics, while V-BLINDS [123], and SSDCT [124] are NR-IQA video quality metrics, which use SVR. CORNIA [116] uses the SVR model to learn from unlabeled data. For example, Freitas *et al.* [125] used the RFR model in his quality model, which is based on ensemble learning that brings together multiple decision tresses [115]. Hui *et al.* [126] used handcrafted spatial and temporal features and the SVM model to blindly predict the quality of the video.

There are several types of neural networks, as shown in Figure 2.8. For quality assessment methods, most commonly used neural networks are multilayer perceptron (MLP), CNN, and Long Short-Term Memory Network (LSTM-N). The multilayer perceptron is an artificial neural network structure that can be used for classification and regression. It is fully connected, which means that every node in a layer is connected to each node in the next layer. For example, in SINGH2019 a multilayer perceptron with a single hidden layer and four neurons has been used [127] to model a NR video quality assessment method. The Long Short-Term Memory is a type of recurrent neural network (RNN) where connections between nodes form a directed graph along a temporal sequence. This allows it to exhibit temporal dynamic behavior. LSTM-N can use their internal state (memory) to process sequences of inputs.

CNN is a well-known deep learning architecture inspired by the natural visual perception mechanisms of living things. In 1990 LeCun[128] published a modern CNN framework and later improved the architecture of the convolutional neural network. However, due to the lack of large training data and computational power, at that time their networks cannot perform well on more complex problems, such as large-scale image and videos. Since 2006, many methods have been developed to overcome the difficulties in training deep CNNs [129]. Most notably, Krizhevsky proposed a classical CNN architecture and showed significant improvements over previous methods in image classification. The overall architecture of their method, AlexNet[130], is similar to LeNet-5 but has a deeper structure. Due to the success of AlexNet, many works

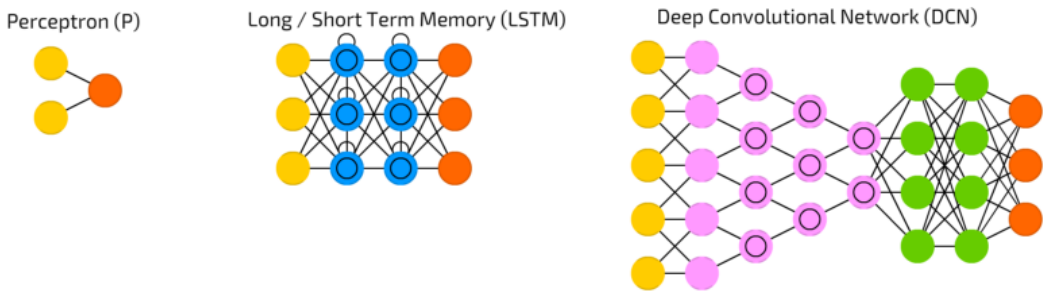

Figure 2.8: Illustration of different types of neural networks.

have been proposed to improve its performance. Among them, four representative works are ZFNet[131], VGGNet[132], GoogleNet [133], and ResNet[134]. A typical trend in the evolution of architectures is that networks are becoming deeper and deeper.

Lee Kang[135] proposed one of the first NR image quality assessment (IQA) (2D images) methods in their work, 32×32 patches are used as input to the CNN architecture. Apart from the input and output layers, in the hidden layer, there is one convolution layer, one pooling layer, and two fully connected layers. Domonkos [136] have developed an no reference video quality assessment method that uses frame level features, which were obtained from a pre trained CNN using transfer learning. A temporal pooling using a regression algorithm (SVR) is used to aggregate these frame-level features for each video and predict overall quality. Singh and Aggarwal [127] have proposed a no reference video quality assessment model in which spatial and temporal features are extracted by a three-dimensional local binary pattern (LBP) operator. These features are mapped to a single scalar quantity using a simple two-layer artificial neural network (ANN) with a single hidden layer of four neurons. Ahn and Lee [137] have proposed an NR Deep Blind Video Quality Assessment (DeepBVQA) method, in which spatial features are extracted by a CNN, named BIECON [138], while temporal features are handcrafted. The final quality score of the video is computed using a feature vector generated by aggregating the pooled frame-level features. Domonkos and Szirányi [139] developed a no reference video quality assessment model based on a long short-term memory (LSTM) network. The method considers the video frames as a time series of deep features, extracted with the help of a CNN, and uses an LSTM network to predict the video quality scores.

2.3.1 Deep Learning Methods

Traditional ML-based algorithms are based on hand-crafted features obtained from a feature engineering task. On the other hand, deep learning-based methods do feature engineering and learn from these features without human intervention, which is why deep learning has become more popular over time. The formulation of neural networks is inspired by the human brain and the human brain consists of billions of neurons interconnected to each other. Each neuron receives the signal, processes it, and passes it on to the other neurons. This is how information is passed on

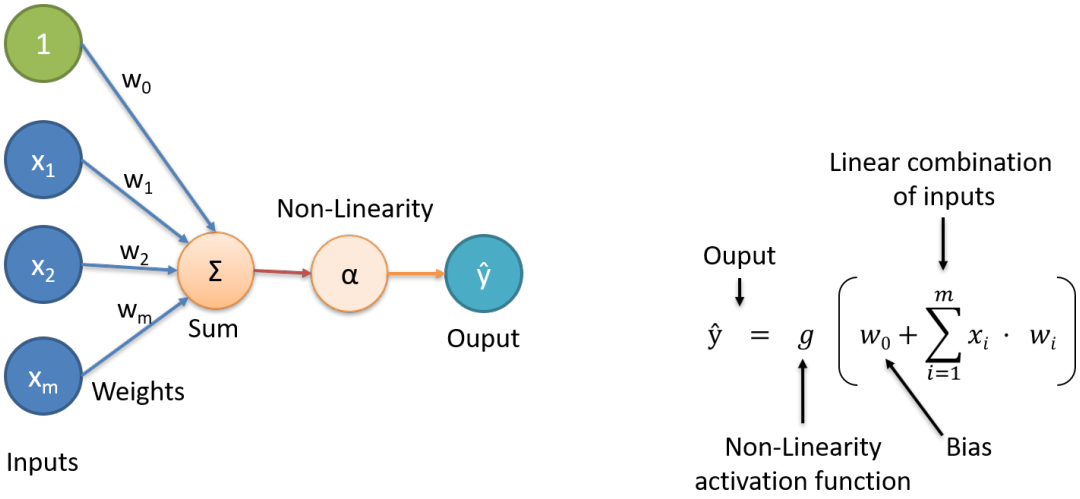

Figure 2.9: A perceptron in forward propagation.

in our brain. Likewise, deep learning focuses on using neural networks to automatically extract patterns in raw data and then using these patterns or features to learn how to perform a task. Traditionally, machine learning algorithms define a set of features in the data. Usually, these features are handcrafted or hand-engineered, and, as a result, they tend to be pretty brittle in practice. The key idea of deep learning is to learn these features directly from data in a hierarchical manner to detect a face, for example, start by detecting the edges in the image, composing these edges together to detect middle-level features, such as an eye or a nose or mouth, and then, going deeper, composing these features into structural or facial features to finally recognize the corresponding face. This hierarchical way of thinking is really the core of deep learning.

An important question arises, "why we are considering deep learning now?". The answer to this question is that the data have become much more pervasive now. Deep learning models are extremely hungry for data, and we are able to get a huge amount of data easily from different online sources. Second, now we have powerful GPU hardware to run deep learning algorithms in parallel processing. And finally, due to open source toolboxes like TensorFlow, Keras, and PyTorch, building and deploying these models has become streamlined.

The fundamental building block of deep learning is a single neuron (also known as a perceptron). As shown in Figure 2.9, a single neuron works in a forward propagation format of the information that passes through it. We can divide a set of inputs x_i to x_m, and each of these inputs or each of these numbers is multiplied by their corresponding weights w_i and then added together. We take this single number, which is the result of addition, and pass it through a non-linear activation function to produce our final output $\hat{y}$. We also have a bias function, which is a shift activation function. The right-hand side of the figure illustrates the forward propagation of a perceptron, and the left-hand side illustrates the mathematical representation of a perceptron. We can re-write this concept in more concise way as follows:

$$\hat{y} = g(w_0 + \mathbf{X}^{\mathbf{T}}\mathbf{W}), \tag{2.21}$$

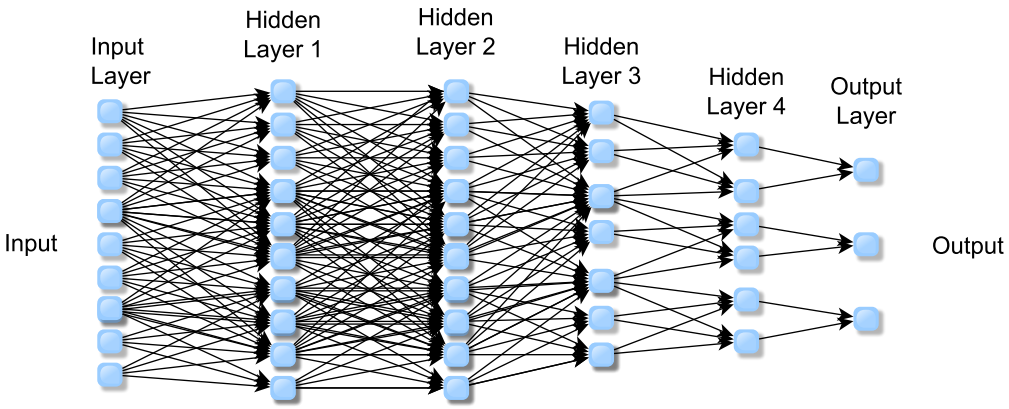

Figure 2.10: Deep Neural Network (DNN)

.

where $\mathbf{X}$ represents a vector of inputs x_1 to x_m at time T, $\mathbf{W}$ represents a vector of weights from w_1 to w_m. The output $\hat{y}$ is simply obtained by taking the dot product of $\mathbf{X}$ and $\mathbf{W}$, adding a bias w_0, and then applying a non-linearity g.

Using one perceptron, we can build a DNN by simply stacking the layers of perceptrons to create more and more hierarchical models, where the final output is computed by going deeper and deeper into the network. Figure 2.10 shows an example of a deep neural network with many hidden layers and many nodes in each hidden layer. Each layer has neurons or perceptrons interconnected with neurons in the next layer. The layer into which we feed the input is called the input layer. The number of nodes in this layer depends on the number of dimensions of the data. The output layer is the layer in which the output is generated. The number of nodes depends on the number of classes in the classification problem or the scalar values in the regression problem. The hidden layer is like the black box in which the extraction of features takes place. The number of hidden nodes and the number of hidden layers are arbitrary. For example, in image classification, every hidden layer extracts features that help identify images. The first hidden layer may extract features, such as edges. The second hidden layer builds upon the features extracted from the first layer and may extract features related to the objects, e.g., the structure of different faces. The more hidden layers we increase, the more complex features extracted [114].

To mathematically represent this deep neural network, first we define the dot product, summation of input vectors, and their corresponding weights:

$$z_{k,i} = w^k_{0,i} + \sum_{j=1}^{n_{k-1}} g(z_{k-1,j}) w^{(k)}_{j,i}, \tag{2.22}$$

where k is the number of layers, n is the number of inputs, $w_{j,i}$ is the i^{th} weight of the perceptron of the j^{th} input, $w^k_{0,i}$ is the bias of the i_{th} input of the k_{th} layer, and z represents the dot product and summation of the input vectors and their corresponding weights before applying the nonlinearity,

and it can be written as follows:

$$z_i = w_{0,i}^k + \sum_{j=1}^{m} x_j w_{j,i}^k. \tag{2.23}$$

Then, we can obtain our output $\hat{y}$ as follows:

$$\hat{y} = g\left(w_{0,i}^k + \sum_{j=1}^{d_k}\right) g(z_j) w_{j,i}^k, \tag{2.24}$$

Where k is the number of layers, $w_{j,i}^k$ is the weight of the j^{th} perceptron of the i^{th} input of the k_{th} layer, z_j is the output of the j^{th} perceptron, d_k represents the desired output value of the perceptron in layer k, and g is a nonlinear activation function.

The nonlinear activation function allows us to deal with nonlinear data because, in the real world, data are always nonlinear. In quality assessment methods, the Exponential Linear Unit (ELU) activation function [140] is commonly used because this function tends to converge faster and produces accurate results. When training a neural network, we want to find a network that minimizes the empirical loss (average loss over the entire dataset) between the predictions and the ground truths (MOS in the case of quality assessment). For quality assessment methods, Mean Squared Error (MSE) loss is a commonly used loss, which can be computed as follows:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2, \tag{2.25}$$

where n is the number of data points, Y_i represents ground truth values and $\hat{Y}_i$ represents predicted values. To find the weights of the neural network, which will minimize the loss of the training dataset, optimization functions are used, such as SGD [141], which is commonly used for loss optimization. In the training process, specifying the learning rates determines how many steps the loss optimization function takes to reach local minima.

For quality assessment methods, the most commonly used deep learning methods are CNN and LSTM. CNN consists of convolutional layers that perform a convolution operation on multi-dimensional input images. Let us take an example of a 2D image. Suppose that we have a 4×4 patch or filter, which will consist of 16 weights. We will apply the same 4×4 filter to the input and use the result of that operation to define the state of the next layer of neurons. Therefore, the neuron in the next layer will be defined by applying this patch with a filter of the same size and learned weights. Then, we are going to shift that patch on the input image by one pixel to get the next patch and compute the next output neuron. This is how the convolution operation works. Using small patches of the image, convolution learns the features of the image while preserving the spatial relationships between pixels.

Each neuron in the CNN layer will compute a weighted sum of each of its patch inputs. We apply and activate the neuron with some nonlinear activation function, so that we can handle nonlinear data relationships. We also need to add a bias in the summation operation that allows

the activation function to be shifted. In other words, we can say that each neuron in the hidden layer only sees a very specific patch of its inputs. It does not see all input neurons. In this case, each neuron output observes only a very local connected patch as input. We take a weighted sum of those patches, apply a bias, and then obtain a feature map (FM) as a result of a convolution layer. The feature map represents the state of the neuron in the next layer. We can define the convolutional layer mathematically as follows:

$$y_{FM} = g\left(\sum_{i=1}^{m}\sum_{j=1}^{n} w_{i,j}x_{(r(i)+p,r(j)+q} + b\right) \tag{2.26}$$

where $w_{i,j}$ represents $i \times j$ filter or patch matrix $i \times j$, $x_{i+p,j+q}$ represents the patch size $p \times q$ in the input image x and r is the dilation rate. Specifically, using this equation, an element-wise multiplication is performed using every element in w by the corresponding elements in the input x. We add bias b and activate it with nonlinearity g. For each neuron in the hidden layer, y_{FM} becomes the input of the neurons in the next layer.

The general layers in DNN-based deep learning methods are as follows [142, 143, 144]:

- Input Layer: The input layer is the layer associated with the input image data. The input layer is usually a tensor with dimensions, such as the dimensions of the input image, namely the length, width, and number of channel images or their transformations.
- Convolution Layer: Convolutional layers are layers that carry out the convolution process from the previous layer. This layer stores the parameters or weights of the training results. The output of this layer (in the form of a tensor, often referred to as a feature map) usually has a length and width smaller than the input layer but a greater depth. The movement of the filter in the image is controlled by the parameter *stride* [144].
- Activation Layer: It is an activation function that decides the final value of a neuron, as described in Eq. 2.21. These functions convert linear input signals into non-linear output signals, which supports the learning of deep networks.
- Pooling Layer: This layer is responsible for reducing the spatial size of the convolved feature. This is to decrease the computational power required to process the data through a dimensionality reduction, but extract dominant features that are rotation and position invariant.
- Fully Connected Layer: It is the final layer that functions as a classifier or a regressor. This layer generally uses artificial neural networks that can be trained. This layer stores the weight of the training results.

3 PERCEPTUAL QUALITY ASSESSMENT OF ENHANCED IMAGES

In this chapter, we present a psychophysical study in which we analyze the perceptual quality of images enhanced with various types of enhancement algorithms. To estimate and compare the quality of the enhanced images, we performed a psychophysical experiment based on the Double Stimulus Continuous Quality Scale (DSCQS) methodology. To perform an online psychophysical experiment, we designed a crowdsourcing interface. We also performed another experiment in a controlled laboratory environment and compared its results with the crowd-sourcing results.

3.1 INTRODUCTION

Our main goal is to introduce a quality database for enhanced images. To our knowledge, currently there is only one image enhancement quality database [145]. However, this database contains only low-resolution images that were manually processed using professional photo editing software (Adobe Photoshop). We present a database containing images with higher resolution that are enhanced with twelve different standalone image enhancement algorithms. Our goal is to produce a set of images similar to what is found in popular consumer applications. Most importantly, the database contains subjective scores for each image, which were obtained with a crowd-sourcing experiment.

In classical image enhancement algorithms, the value of the pixels is manipulated to achieve the desired enhancement effect. For example, the unsharp mask filter is a simple and highly versatile sharpening algorithm that enhances fine details by adding high-frequency spatial information to the original image [146]. An unsharp mask is simply an out-of-focus image created by spatially filtering the sample image with a Gaussian low-pass filter. This filter can be viewed as a convolution operation of an image with a kernel mask that is a two-dimensional Gaussian function (g(x,y)). A Gaussian lowpass filter is used to blur the image and remove noise [147].

Histogram equalization algorithms (HE) are very popular contrast enhancement methods [89], which are used in many applications, such as medical image enhancement and underwater image enhancement. There are also other variations of the HE algorithm that can be used for image enhancement. For example, the dynamic histogram equalization (DHE) algorithm [148] partitions the input histogram into sub-histograms so that there is no dominating component. Each sub-histogram goes through HE, providing a better overall contrast enhancement because of the controlled dynamic range of the gray levels, which eliminates the compression of low histogram components. Another algorithm is the average histogram equalization (AHE), which is an excellent contrast enhancement method for natural and medical images. Another variant of HE is the

Hue Saturation Histogram Equalization (HSHE) [149], which mirrors the properties of the global HE method in the local HE method to avoid artifacts while improving global and local contrasts. The HSHE algorithm can be applied in two ways: by processing the luminance or by processing each channel individually.

Another type of enhancement algorithm targets low-light scenarios, which often contain a lot of noise. Ren *et al.* proposed a joint low-light enhancement and noise reduction strategy [150] that aims to enhance low-light images, mitigating the inherent noise problem. This method performs a decomposition based on the Retinex model into successive sequences, which sequentially estimates a piecewise-smoothed illumination and a noise-suppressed reflectance. After the illumination and reflectance map are obtained, the illumination level is adjusted to generate the enhanced image. In this work, we use the above-cited enhancement algorithms to produce images with different quality levels.

3.2 CROWDSOURCING EXPERIMENTS

As mentioned earlier, subjective experiments are considered the ground-truth in image quality. To guarantee the validity, reliability, and reproducibility of subjective quality assessment methods, over the years, several recommendations have been drafted for experimental methodologies. Examples of popular recommendations are Rec. ITU-R BT.500 [74] and Rec. ITU-T P.910. [73]. It is worth mentioning that a good number of these experimental methodologies have been derived from classic psychometric practices [151]. Experimental methodologies have different specifications, including the presentation of stimuli, the type of subjective scale, and the scoring procedure. Each type of methodology has advantages and disadvantages, and it is difficult to cover all factors to provide a specific set of recommendations. Pinson *et al.* [152] have detailed several aspects that should be taken into consideration when performing an experiment, such as the environment conditions, the number of participants and stimuli, and the scoring procedure.

In terms of the subjective scoring procedure, methodologies can be divided into rating- or ranking-based methods [152]. In ranking-based methods, participants are asked to rank images in terms of their quality or to compare each pair of images of the set (a pairwise comparison, PC) [152]. For example, PC methodologies require that participants compare all possible combinations of image pairs, which is a very time-consuming process. PC methodologies are often considered when the differences between stimuli are small and, therefore, hard to differentiate. In rating-based methods, participants assign a score to each stimulus presented to them, using either a numerical scale or a category scale. These methods are generally less time-consuming than rank-based methodologies and, consequently, more popular. For all rating-based methods, the final subjective score of a stimuli, i.e., the MOS (mean opinion score), is computed by taking the average of the scores over all subjects.

In terms of stimuli presentation, rating-based methodologies can be single-stimulus (SS),

double-stimulus (DS), or multi-stimulus (MS). In SS methodologies, participants rate the quality of just one stimulus (the test), without having a reference. In DS and MS methodologies, participants rate the quality of two or more stimuli, which are presented simultaneously or closely spaced in time. According to the rating scale used, the DS methodologies can be classified as Double Stimulus Continuous Quality Scale (DSCQS) [74] or Double Stimulus Impairment Scale (DSIS) [153]. In DSIS (also referred to as Degradation Category Rating - DCR), participants rate both displayed stimuli using a discrete 5-point impairment scale: Imperceptible (5), perceptible but not annoying (4), slightly annoying (3), annoying (2), very annoying (1). In DSCQS, participants rate the quality of both the reference image and the test image using a 5-point quality scale: Excellent (5), Good (4), Fair (3), Poor (2), and Bad (1). For SS methodologies, a popular scale is the absolute category rating (ACR) [73], in which volunteers rate images using a 5-point impairment scale or a 5-point quality scale (as described above).

In summary, traditional subjective experiments require the following:

1. a diversified pool of participants;
2. a laboratory environment, with physical conditions that adhere to the recommendations (e.g., Rec. ITU-R BT.500 [74] and Rec. ITU-T P.910. [73]);
3. a dataset of test stimuli that the participants will assess.

In recent years, crowd-sourcing experiments have been used in various application areas as a cost-effective substitute for subjective experiments in a laboratory setting [59, 75, 76, 77, 78, 79]. In other words, instead of conducting the subjective experiment in a controlled laboratory setting, researchers use dedicated platforms (e.g., Amazon Mechanical Turk[1] and Microworkers[2]), social networks (Twitter, Facebook, Linkedin), or email campaigns to recruit subjects to participate remotely in subjective studies. Crowdsourcing experiments have many advantages, including the ability to recruit large numbers of participants. Additionally, crowdsourcing campaigns typically require a short implementation time and can collect data from a variety of subjects with different backgrounds, testing environments, and devices.

Although crowd-sourcing platforms help collect data from a larger population, participants are unsupervised and evaluate stimuli at their leisure using their own hardware. In other words, experimenters have limited control over how and when participants perform the experimental tasks, such as the duration of the experimental session, the lighting in the room, and the displays / devices used by participants, which are all important factors that influence the perceived quality. For example, in laboratory experiments, a fixed amount of time is allocated for the experiment, which is often broken down into sessions to avoid fatigue. However, in crowdsourcing, participants perform the experiment at their convenience, taking pauses or interrupting the session when they feel like. In fact, it is known that crowdsourcing participants are reluctant to do longer sessions [78]. Finally, compared to laboratory experiments, crowdsourcing experiments have a

[1] https://www.mturk.com
[2] https://microworkers.com

higher number of unreliable participants and, therefore, a certain number of unreliable quality scores [69]. However, when performed carefully, crowdsourcing experiments may yield results similar to laboratory experiments [78]. In this work, we performed a crowdsourcing subjective experiment using the DSCQS methodology.

3.3 DATABASE CONTENT GENERATION

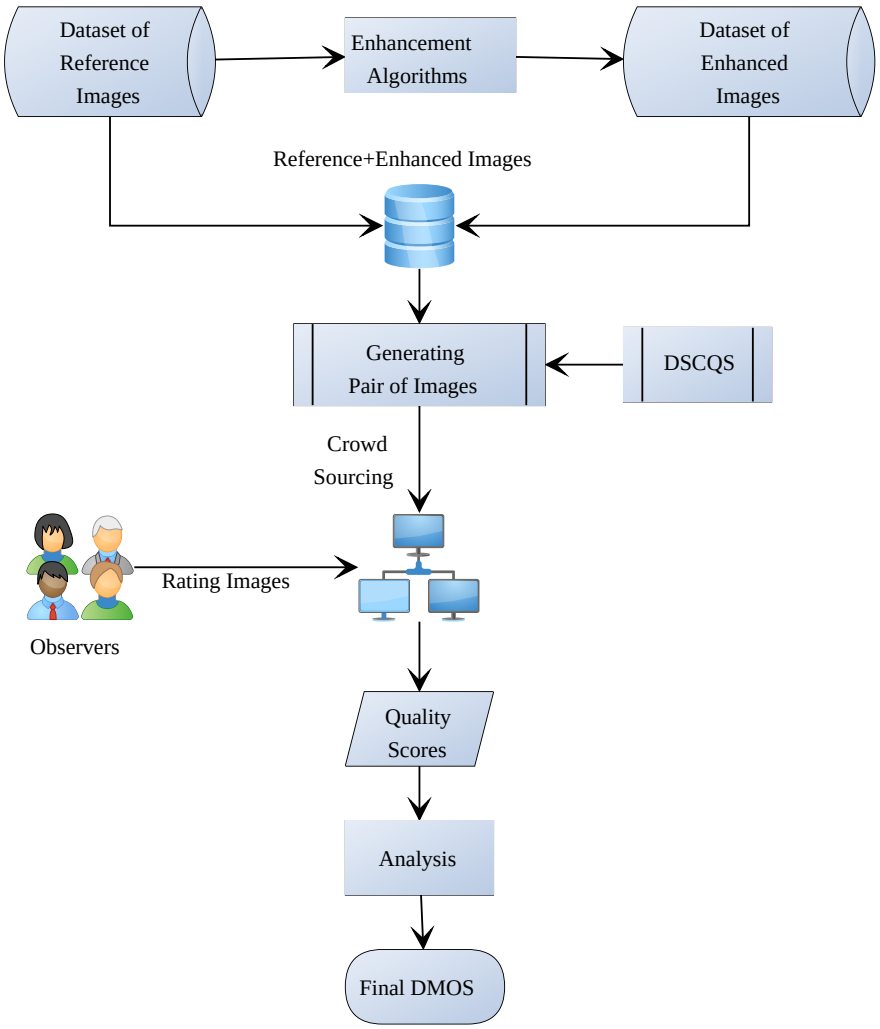

Figure 3.1: Block diagram of the strategy used to create the database and run the crowd-sourcing experiment.

Figure 3.1 shows a block diagram of the strategy used to generate the database. Our first step was to choose 35 original (source - SRC) images. These images were taken from three image quality databases to allow for future comparisons of enhanced and degraded images. More specifically, we took 5 SRC images from the CSIQ database [5], 5 original images from the TID2013 database [4], and 25 original images from the ChallengeDB database [3]. Table 3.1 shows a list of SRC images, along with their names in the corresponding databases. These chosen source

Table 3.1: List of SRC images of the experiment, which were taken from the Challenge database [3], TID2013 [4] and CSIQ [5] databases.

SRC	Database Name	Name in Databases
SRC01	Challenge database	10
SRC02	Challenge database	17
SRC03	Challenge database	129
SRC04	Challenge database	50
SRC05	Challenge database	113
SRC06	Challenge database	138
SRC07	Challenge database	147
SRC08	Challenge database	152
SRC09	Challenge database	167
SRC10	Challenge database	173
SRC11	Challenge database	186
SRC12	Challenge database	255
SRC13	Challenge database	261
SRC14	Challenge database	271
SRC15	Challenge database	338
SRC16	Challenge database	344
SRC17	Challenge database	414
SRC18	Challenge database	442
SRC19	Challenge database	444
SRC20	Challenge database	452
SRC21	Challenge database	455
SRC22	Challenge database	500
SRC23	Challenge database	525
SRC24	Challenge database	527
SRC25	Challenge database	820
SRC26	CSIQ database	1600
SRC27	CSIQ database	boston
SRC28	CSIQ database	child swimming
SRC29	CSIQ database	trolley
SRC30	CSIQ database	woman
SRC31	TID2013 database	103
SRC32	TID2013 database	104
SRC33	TID2013 database	107
SRC34	TID2013 database	111
SRC35	TID2013 database	123

contents are diverse in terms of spatial activity, semantic content, and color distribution. The first row (SRCs) in Figure 3.2 shows examples of SRC images taken from the (a-b) TID2013, (c-d) CSIQ, and (e-f) ChallengeDB databases.

Our next step consists of choosing the enhancement algorithms to be used as different test conditions of our experiment. In this thesis, we refer to these test conditions as Hypothetical Reference Circuits (HRC). We chose a total of 12 enhancement algorithms: Color Enhancement (CE), Contrast Enhancement (CrE), Sharpness Enhancement (SE), Brightness Enhancement (BE), Unsharp Masking (UM), Gaussian Blur (GB), Histogram Equalization (HE), Dynamic His-

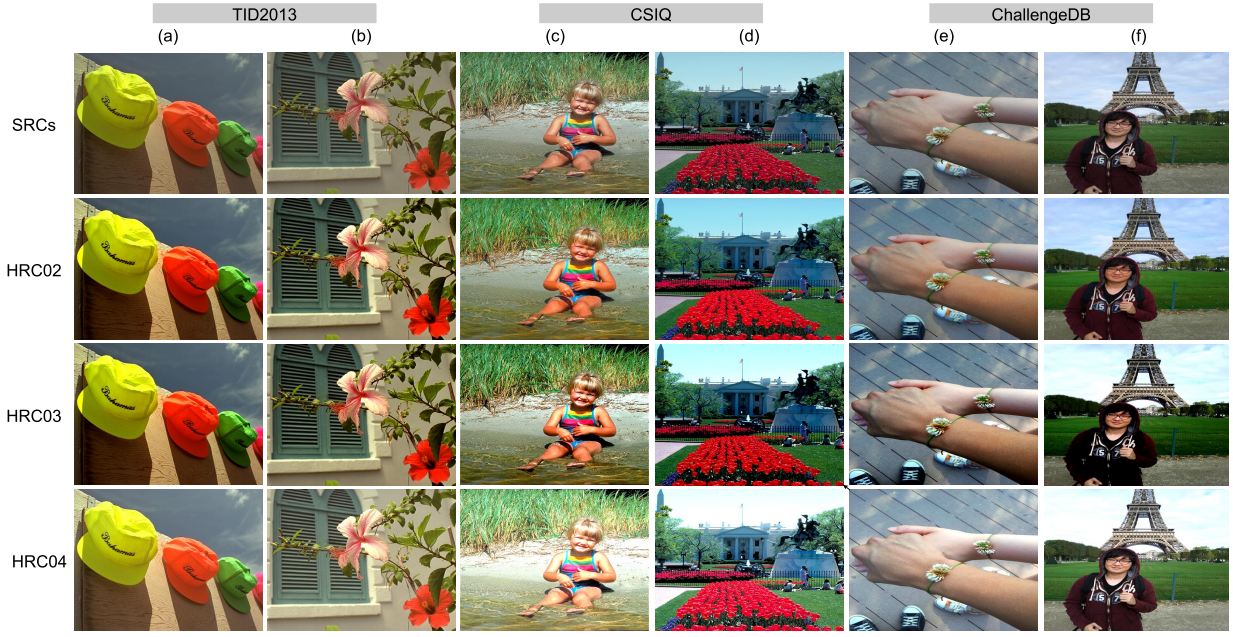

Figure 3.2: Sample source (SRC) images used in our database, processed with different enhancement algorithms (HRCs - see Table 3.2). SRC images were taken from the (a-b) TID2013, (c-d) CSIQ, and (e-f) ChallengeDB databases.

togram Equalization (DHE) [148], Exposure Fusion Framework (EFF) [154], Average Histogram Equalization (AHE), Hue Saturation Histogram Equalization (HSHE), and Joint Enhancement and Denoising Method via Sequential Decomposition (JEDMSD) [150]. Table 3.2 shows a list of the HRCs of the database and the corresponding enhancement algorithms. We use traditional versions of these algorithms, which were implemented using *Pillow package* [155] in Python [156] and Matlab [157]. To generate the test images, we processed each of the 35 SRCs using all HRCs, generating a total of 420 enhanced images. The last three rows in figure 3.2 list the images processed with HRC02, HRC03, and HRC04, which were generated from the SRC images shown in the top row.

Given the number of test images (420), we manually divided the set of stimuli into 3 subsets. Each subset contains 140 images, with different combinations (non-factorial) of at least 10 HRCs and 35 SRCs. To avoid presentation bias, we generated 4 versions of these 3 subsets, which had different HRC-SRC combinations. In total, there were 12 groups of different test images. In this way, each participant did not rate all possible combinations of HRCs and SRCs, but only a third of them, which made the experiment less tiring.

3.4 EXPERIMENTAL METHODOLOGY

Each experimental session is divided into four stages. In the first stage, called *the registration*, the participant fills out a form with personal information. In the second stage, called *the training*, the participant watches a set of sample images and their corresponding enhanced versions. The

Table 3.2: Enhancement algorithms and their corresponding Hypothetical Reference Circuits (HRC) in the experiment.

HRC	**Enhancement Algorithm**	**Abbreviation**
HRC01	Color Enhancement	CE
HRC02	Contrast Enhancement	CrE
HRC03	Brightness Enhancement	BE
HRC04	Sharpness Enhancement	SE
HRC05	Unsharp Masking	UM
HRC06	Gaussian Blur	GB
HRC07	Histogram Equalization	HE
HRC08	Dynamic Histogram Equalization	DHE
HRC09	Exposure Fusion Framework	EFF
HRC10	Average Histogram Equalization	AHE
HRC11	Hue Saturation Histogram Equalization	HSHE
HRC12	Joint Enhancement and Denoising Method via Sequential Decomposition	JEDMSD

goal is that the participant familiarizes himself/herself with the quality range of the images in the database. In the third stage, called *the practice*, the participant performs the scoring procedure by performing a small number of trials, identical to those in the experimental session. Finally, in *main experimental session*, the participant rates the quality of an SRC image and its enhanced version, which was processed with a specific HRC. Figure 3.3 shows a screenshot of the interface, showing this scoring procedure for each trial. In the interface, the 5-point quality scales are shown below each image. The positions (left or right) of the SRC and its enhanced version are randomized for each trial. The database created in this work is available for download on the GPDS site[3].

3.5 CROWD-SOURCING EXPERIMENTAL RESULTS

A total of 108 participants participated in the crowd-sourcing experiment, 78% of the participants being male and 22% female. The participants' age ranged from 17 to 63 years. To analyze the gathered data, we computed the mean observer scores (MOS) and the difference MOS (DMOS). MOS is calculated by averaging the scores given by all participants for each HRC and each test image. DMOS is computed by taking the average of the differences between the scores given to a test image and the score given to its corresponding SRC. In other words, we average the differences between the scores of the two images, which are shown jointly to participants in each experiment trial.

Figure 3.4 presents the DMOS for each of the HRCs and SRCs. Each plot in this figure corresponds to the DMOS of a single HRC across the different SRCs (x-axis) from this figure it is hard to identify any patterns from the graphs. To take a closer look at the results and minimize

[3] <http://www.ene.unb.br/mylene/databases.html>

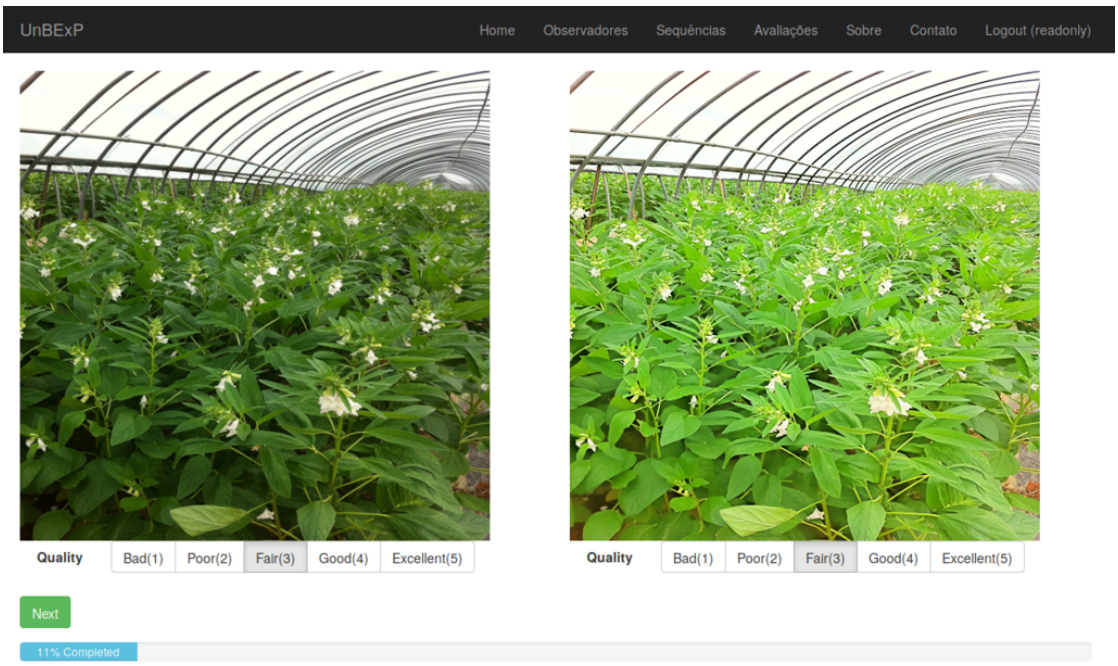

Figure 3.3: The crowd-sourcing experimental interface, displaying an SRC image and its version processed with a specific HRC.

the effect of image content, we compute the average of both DMOS and MOS values for each HRC across all SRCs. Figures 3.5 and 3.6 show the mean MOS and DMOS values, respectively, for each individual HRC, along with their confidence intervals. For DMOS, negative values indicate that the enhancement algorithm (on average) reduced the perceived quality of the SRC image, while positive values indicate that the algorithm (on average) improved the quality of the SRC image. Notice that HRC01, HRC04, and HRC05 produced positive average DMOS values, while HRC06-HRC12 produced negative average DMOS values. For brightness (HRC02) and contrast enhancement algorithms (HRC03), the results were inconclusive. The Sharpness Enhancement algorithm (HRC04) produced the highest DMOS values, followed by the Unsharp Masking (HRC05) and Color Enhancement (HRC01) algorithms.

To verify if the participants perceived differences on the algorithm's image quality when comparing two HRCs, we executed a paired-samples t-test. This test checks if there is a statistically significant difference between the average DMOS considering pairs of HRCs. In each pairwise comparison, we considered only the cases where the participant scored both HRCs. The most relevant results of this test are shown in Table 3.3. Notice that the average DMOS of six HRC pairs is not statistically significant ($p > 0.05$): HRC01-HRC05, HRC03-HRC10, HRC07-HRC08, HRC07-HRC09, HRC08-HRC09, HRC11-HRC12. This means that participants (on average) did not see a difference in quality between these pairs of HRCs. For all other combinations of HRC pairs, the test found statistical differences between the average DMOS, which means that participants (on average) can distinguish the image quality produced by the remaining combinations.

To verify whether content (SRC) affects quality perception, we performed a one-way ANOVA to determine if there is a difference that is statistically significant between average DMOS, consi-

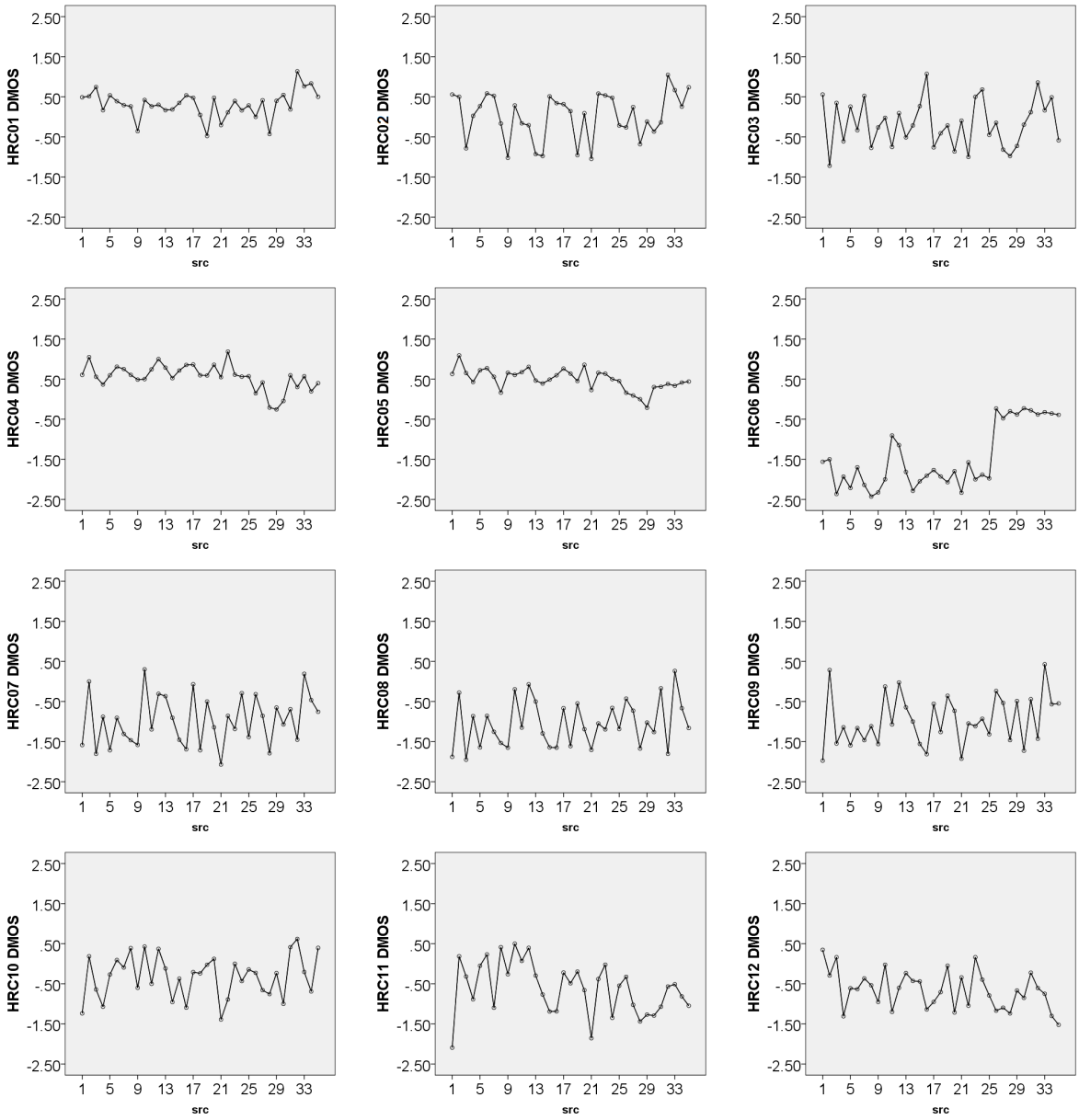

Figure 3.4: Average DMOS values versus the SRC image, for each HRC.

Table 3.3: Paired sample t-test pairs for which the differences in DMOS were not statistically significant.

Pair	HRC	N	Mean	Std.dev	t	p
HRC01-HRC05	HRC01	222	0.4640	1.0746	-1.562	0.120
	HRC05	222	0.5946	0.9643		
HRC03-HRC10	HRC03	293	-0.1640	1.2165	0.838	0.403
	HRC10	293	-0.2389	1.3564		
HRC07-HRC08	HRC07	578	-0.9498	1.4431	0.795	0.427
	HRC08	578	-0.9896	1.3932		
HRC07-HRC09	HRC07	588	-0.9388	1.4568	0.836	0.404
	HRC09	588	-0.9796	1.4296		
HRC08-HRC09	HRC08	590	-1.0186	1.3856	-0.071	0.944
	HRC09	590	-1.0153	1.4195		
HRC11-HRC12	HRC11	624	present -0.5048	1.4255	1.601	0.110
	HRC12	624	-0.6154	1.4532		

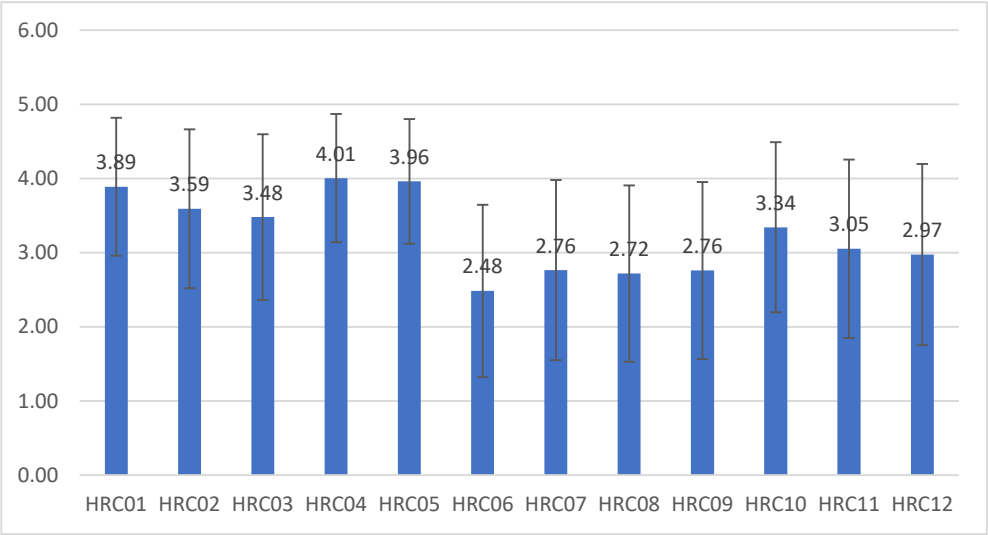

Figure 3.5: Mean Observer Score (MOS) computed across all SRCs for each HRC (see Table 3.2).

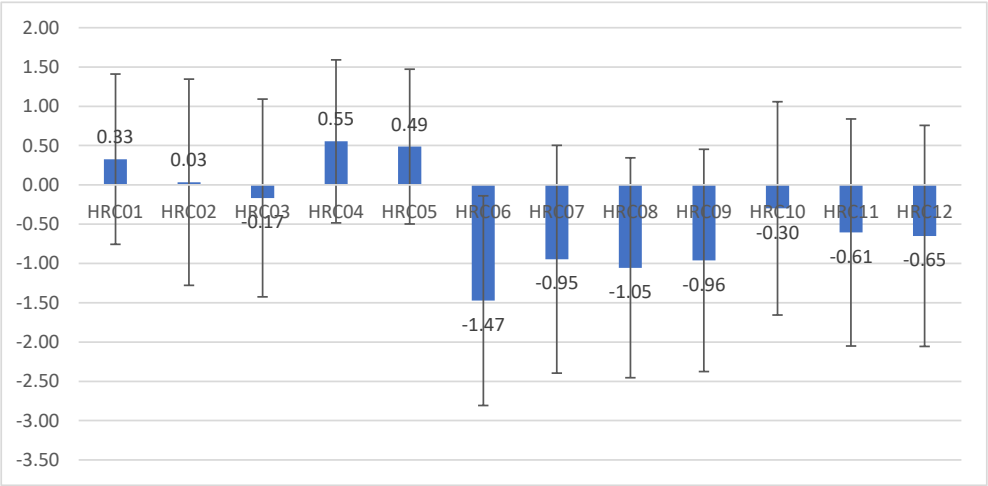

Figure 3.6: Difference Mean Observer Score (DMOS) computed across all SRCs for each HRC (see Table 3.2).

dering SRC as an additional factor. For all groups (HRCs), ANOVA returned a value of $p < 0.05$, meaning that there is at least one pair of SRCs, considering each HRC, that has a difference in the average DMOS that is statistically significant. To identify these pairwise comparisons, we performed a Tukey-Kramer post-hoc test, which allows different group sizes, but gives the same result as the Tukey post hoc test would give if the group sizes were equal. The test shows that 1,151 combinations (out of 7,141 possible combinations) have differences in average DMOS that are statistically significant. This means that in 16% of the cases, the participants found that an enhancement algorithm led to a visible difference for a specific pair of SRCs. We see this behavior by observing the DMOS values in Figure 3.4, which change according to the SCRs.

Another output parameter of the Tukey-Kramer test is the number of homogeneous sets. Table 3.4 shows these results, where SCRs are classified as having one or more sets, depending on their distance to the mean. In a scenario where there is no statistically significant difference between the means, each HRC would have only one homogeneous set that contains all SRCs. In other

Table 3.4: Number of Homogeneous Sets found by the Tukey-Kramer post-hoc test.

HRC	Number of Sets
HRC01	6
HRC02	8
HRC03	13
HRC04	7
HRC05	4
HRC06	6
HRC07	11
HRC08	10
HRC09	11
HRC10	12
HRC11	12
HRC12	6

words, for this HRC, the results would be content-independent. If the number of homogeneous sets is small and the sets are disjoint, we can extract the features of these groups that influence the quality differences. Notice that HRC05 has fewer homogeneous sets (4) than the other HRCs. HRC06 is the only HRC that has disjoint sets. But since most SCRs have more than one set, it is not easy to identify a set of features that can distinguish the differences. It is necessary to conduct further research to better understand how content (SRC) affects perceived quality.

3.6 RESULTS OF LABORATORY EXPERIMENTS

Since the reliability of crowd-sourcing methodologies remains questionable [158], we have also carried out an experiment in a controlled laboratory environment and compared the results in the two experiments. Eighteen participants took part in the experiment, 72% being male and 28% female. The age range of the participant ranged from 18 to 70 years. We used the same experimental methodology and interface platform used in the crowd-sourcing experiment. The experimental setup followed the ITU-R BT.500 recommendation [74]. Table 3.5 shows the technical specifications used in the experiment. An experimental session lasted, on average, 40 minutes.

A one-way analysis of variance (ANOVA) was used to compare the DMOS values of these two groups of experimental results: the onsite (laboratory) and online (crowd-sourcing) experiments. Table 3.6 shows these ANOVA results. The last column of the table indicates whether there is a statistically significant difference between the two groups. More specifically, if $p < 0.05$ the differences are statistically significant, and therefore the participants in the two experiments rated the images enhanced with a particular HRC differently. We can see that for HRC07, HRC09, and HRC12 participants in the two experiments rated the images differently, while for the other 9 HRCs, the differences are not statistically significant. Therefore, for most HRCs, there is no statistically significant difference between the DMOS given by the participants in both experiments.

Table 3.5: Technical specifications of the equipment used for the Laboratory Experiment.

Item	**Specification**
Monitor	BENQ XL2420z 1920x1080 144hz
Distance of the observer	60 CM
GPU	QUADRO K4000
Brightness	100%
Sharpness	50%

Table 3.6: ANOVA results comparing online and on-site groups.

HRC	**Group**	**N**	**Mean**	**Std.dev**	**F**	**p**
HRC01	OnSite	193	3.891	0.840	0.002	0.961
	Online	1,255	3.888	.9430		
HRC02	OnSite	196	3.515	.9844	1.344	0.247
	Online	1,250	3.611	1.090		
HRC03	OnSite	197	3.467	1.003	0.051	0.821
	Online	1,250	3.486	1.138		
HRC04	OnSite	197	4.015	0.854	0.060	0.806
	Online	1,248	4.031	.849		
HRC05	OnSite	197	4.015	.785	0.182	0.669
	Online	1,249	3.988	0.839		
HRC06	OnSite	201	2.408	1.146	1.525	0.217
	Online	1,246	2.518	1.172		
HRC07	OnSite	203	2.586	1.060	5.072	0.024
	Online	1,244	2.795	1.2495		
HRC08	OnSite	202	2.629	1.109	1.560	0.212
	Online	1,246	2.742	1.204		
HRC09	OnSite	203	2.542	1.044	8.176	0.004
	Online	1,249	2.801	1.219		
HRC10	OnSite	201	3.269	1.076	1.125	0.289
	Online	1,247	3.362	1.166		
HRC11	OnSite	196	2.913	1.131	2.990	0.084
	Online	1,246	3.074	1.221		
HRC12	OnSite	194	2.778	1.100	5.890	0.015
	Online	1,252	3.008	1.245		

3.7 CONCLUSIONS

In this work, we have built an image enhancement quality database, which is large and diverse. We present a psychophysical study in which we analyze the perceptual quality of images enhanced with different types of enhancement algorithms, including color, sharpness, histogram, and contrast enhancement. A crowd-sourcing (online) psychophysical experiment was performed to obtain quality scores for these images. We also conducted a controlled laboratory experiment (onsite) and compared its results with the crowdsourcing (online) experiment. This database can be used to train image quality metrics that can detect both increases and decreases in perceived quality.

4 NO-REFERENCE UNDERWATER IMAGE QUALITY ASSESSMENT METRICS

In this chapter, we present two quality metrics for underwater images. The first method uses an adapted version of the multi-scale salient local binary pattern operator to extract image features and a machine learning approach to predict quality. This method can be used to evaluate the results of restoration techniques quickly and efficiently, opening up a new perspective in the area of underwater image restoration and quality assessment. The second method is based on a convolutional neural network. We train a CNN model using patches of underwater images. we choose patches that are most perceptually relevant by incorporating the properties of the human visual system, such as visual saliency. These selected patches are then fed to the CNN model for quality estimation.

4.1 NR-UWIQA METRIC BASED ON MULTISCALE SALIENT LOCAL BINARY PATTERNS OPERATORS

The first NR-UWIQA metric is based on a Multiscale Salient Local Binary Patterns (MSLBP) operator [159]. More specifically, this method uses a variant of the local binary pattern (LBP) [160], to extract features that are relevant to image quality. The LBP operator can be computed using the following equation:

$$L_R^P(\mathcal{I}_c) = \sum_{p=0}^{P-1} \sigma(\mathcal{I}_p, \mathcal{I}_c) 2^p, \tag{4.1}$$

where $\mathcal{I}$ is an input image, $\mathcal{I}_c = \mathcal{I}(x, y)$ is an arbitrary central pixel at position (x, y), $\mathcal{I}_p = \mathcal{I}(x_p, y_p)$ is a neighboring pixel surrounding $\mathcal{I}_c$, and $\sigma(u, v)$ is the step function given by:

$$\sigma(v, u) = \begin{cases} 1, & \text{if } v - u \geq 0, \\ 0, & \text{otherwise.} \end{cases} \tag{4.2}$$

The points (x, y) are related to the neighboring points (x_p, y_p) as follows:

$$x_p = x + R\cos\left(2\pi\frac{p}{P}\right) \qquad \text{and} \qquad y_p = y - R\sin\left(2\pi\frac{p}{P}\right).$$

In the above equations, $p = \{1, 2, \cdots, P\}$ is the number of neighboring pixels sampled from a distance of R from $\mathcal{I}_c$ to $\mathcal{I}_p$. Figure 4.1 describes the steps to apply the LBP operator on a single pixel ($\mathcal{I}_c = 8$) located in the center of an image block 3×3, as shown in the bottom left of this figure. The numbers in the gray squares of the block represent the order in which the operator is computed (clockwise direction, starting from 0). In this figure, we use a unitary neighborhood

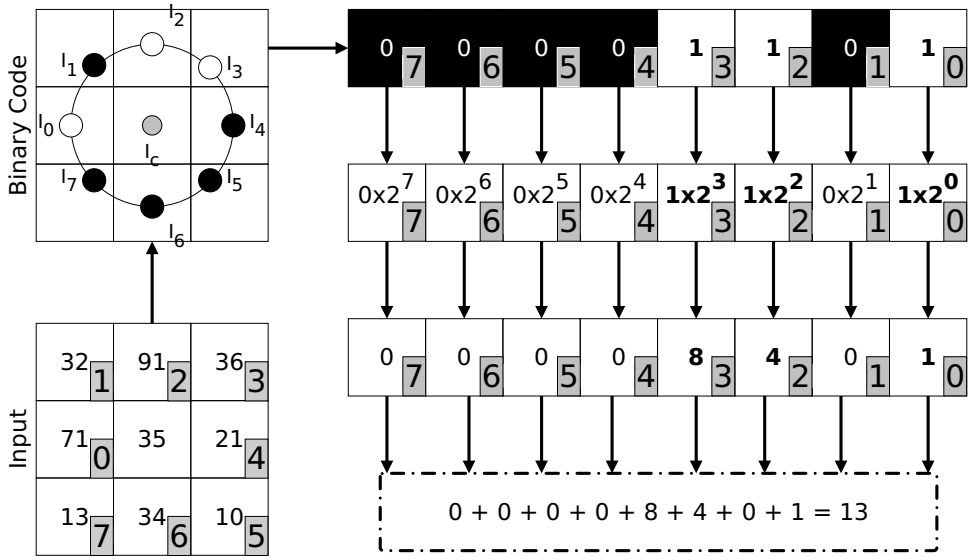

Figure 4.1: Example of LBP algorithm using $R = 1$, $P = 8$, $\mathcal{I}_c = 35$, $\mathcal{I}_p = \{71, 32, 91, 103, 21, 10, 34, 13\}$, and $L_1^8(35) = 13$ [2].

radius ($R = 1$) and eight neighboring pixels ($P = 8$).

After calculating $\sigma(v, u)$ for each neighboring pixel $\mathcal{I}_p$, we obtain a binary output for each $\mathcal{I}_p$ ($0 \leq p \leq 7$), as illustrated in the block in the upper-left position of Figure 4.1. In this block, the black circles correspond to "0" and the white circles to "1". These binary outputs are stored in binary format according to their position (gray squares). Then, the resulting binary number is converted to decimal format. This decimal number is the output produced by LBP for $\mathcal{I}_c$. Then, we compute the LBP labels for all pixels of an image, obtaining the LBP maps. Instead of using single values for R, P, MLBP generates multiscale LBP maps by varying the parameters R and P and performing a symmetrical sampling. For a set of parameters R and P, the MLBP operator computes the LBP labels of all pixels in an image and obtains a set of LBP maps ($\mathcal{L}_R^P$). In MSLBP, the spatial features extracted by the MLBP operator are weighted by a saliency map generated by a Boolean map saliency model (BMS) [161]. BMS saliency maps $\mathcal{W}(x, y)$ have values between 1 and 0, which represent the saliency value of the corresponding pixel in the underwater image. We name each weighted map the salient local binary pattern (SLBP) map, while the weighted maps of the MLBP maps are named multiscale SLBP (MSLBP) maps. The weighted features computed for the MSLBP are used as input to a supervised machine learning algorithm that predicts the final image quality score.

In this work, instead of using the BMS algorithm, we use the Graph-Based Visual Saliency (GBVS) [93] model to generate the saliency maps $\mathcal{W}$. Figure 4.2 shows samples of saliency maps generated by the GBVS model using a few underwater images as input images. We chose the GBVS model because it is a traditional model that is easy to execute and has a performance similar to the BMS model proposed in the original MSLBP method [159]. The saliency maps $\mathcal{W}$ are used to weigh each pixel of the MLBP maps $\mathcal{L}_R^P$. A feature vector is obtained by computing the histogram of the $\mathcal{L}_R^P$ maps weighted by the salience maps $\mathcal{W}$. In particular, the histogram is

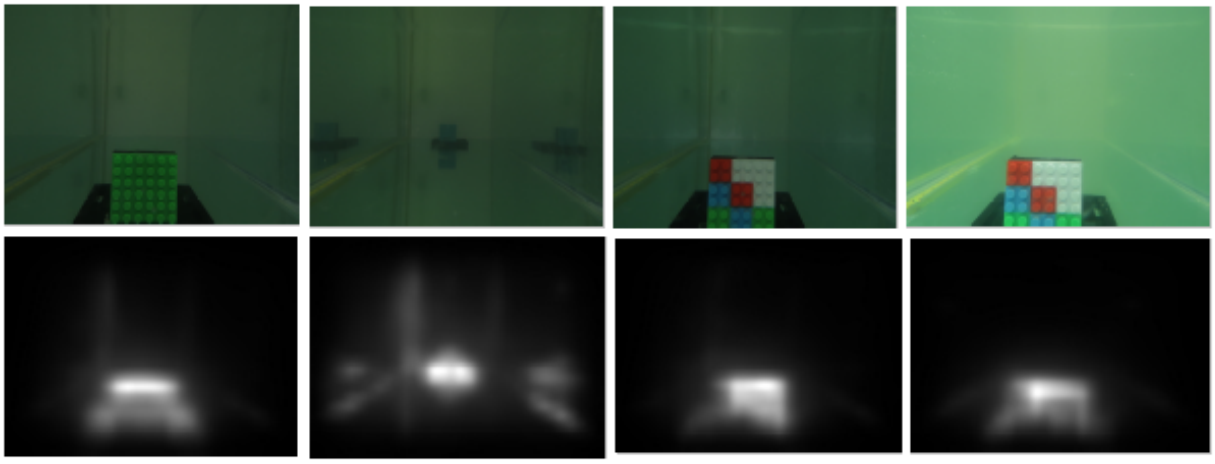

Figure 4.2: Underwater images and their saliency maps.

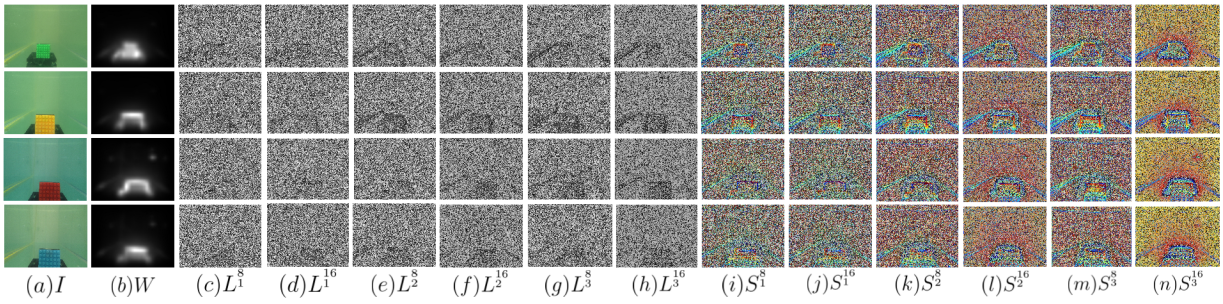

Figure 4.3: Example of underwater images (a), their saliency maps (b), LBP maps (c)-(h), and SLBP maps (i)-(n).

generated as

$$H_R^P = \{h_R^P(0), h_R^P(1), \cdots, h_R^P(P+1)\} \tag{4.3}$$

where

$$h_R^P(\phi) = \sum_{x,y} \mathcal{W}(x,y) \cdot \delta(\mathcal{L}_R^P(x,y), \phi), \tag{4.4}$$

and

$$\delta(v,u) = \begin{cases} 1, & \text{if } v = u, \\ 0, & \text{otherwise.} \end{cases} \tag{4.5}$$

The number of bins of this histogram is similar to the number of different labels in $\mathcal{L}_R^P$. So each $\mathcal{L}_R^P(i,j)$ can be represented by its weighted form, generating the map $\mathcal{S}_R^P$. Figures 4.3 (a) and (b) depict the examples of the input images and their saliency maps, respectively. Figures 4.3 (c) to (h) depict examples of LBP maps obtained using different radius values (R) and different numbers of neighboring points (P). Figures 4.3 (i) to (n) show the SLBP maps generated from $\mathcal{W}$ and their corresponding $\mathcal{L}_R^P$. In this work, we have used $R = 1, 2$, and 3 and $P = 4, 8$, and 16. After generating the SLBP maps, we compute the different SLBP histograms H, as illustrated in Figure 4.4. These histograms are concatenated to produce a feature vector for each underwater

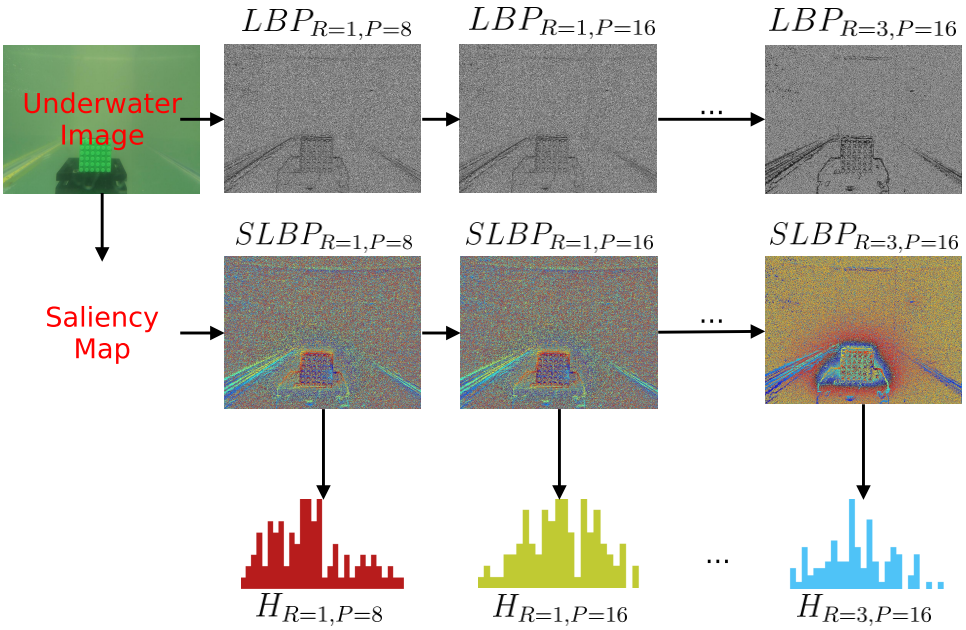

Figure 4.4: Multiple histogram generation from SLBP.

image as follows:

$$\mathcal{H} = H_1^4 \oplus H_1^8 \oplus H_2^4 \oplus H_2^8 \oplus H_2^{16} \oplus \cdots \oplus H_R^N, \tag{4.6}$$

where $\oplus$ denotes the concatenation operator.

The computed feature vector $\mathcal{H}$ is supplied as input to a random forests (RFR) regression algorithm to predict the quality of underwater images. We chose RFR because in previous studies it has shown robust performance values [162] when compared to other machine learning algorithms (neural networks, support vector machines, generalized linear models, etc.).

We used the UID-LEIA database for training and testing the proposed underwater IQA metric. We used the Spearman's Rank Order Correlation Coefficient (SROCC), Pearson's Linear Correlation Coefficient (PLCC), the Kendall Rank Correlation (KRCC), and Root Mean Square Error (RMSE) as performance metrics. We compare the proposed methods with the following publicly available underwater IQA methods: the Underwater Color Image Quality Evaluation (UCIQE) [56] and the Patch-Based Mean Underwater Image Quality (PUIQ) [57]. Additionally, we also compare our method with traditional NR-IQA methods: CORNIA [116], BRIQUE [119], SSEQ [118], DIIVINE [121], NIQE [163], Choi *et al.* [164], and Balboa *et al.* [165]. The experiments were carried out on a computer with an Intel Core i7-4790 processor at 3.60 GHz, running an Ubuntu operating system. From the UID-LEIA dataset, 80% of the content is used for training and 20% for testing. This 80-20 training-testing random split is performed 1,000 times, and then the mean correlation is computed and reported.

Table 4.2 illustrates the performance comparison of the proposed method with other state-of-the-art metrics on the UID-LEIA dataset. Notice that the proposed metric outperforms all other metrics with respect to SROCC, PLCC, and RMSE values. The highest KRCC values are

obtained using the NIQE metric. The values in bold correspond to the best performance, while the results in italics correspond to the second-best results. These results show the advantage of the proposed approach in terms of accuracy.

Table 4.1: Performance evaluation of MSLBP metric with different IQA methods on UID-LEIA dataset.

Type	Method	KRCC	PLCC	SROCC	RMSE
NR-IQA Methods	CORNIA	0.6502	0.4549	0.6394	32.1954
	SSEQ	0.0247	0.0129	0.0199	35.7431
	BRISQUE	0.1719	0.0998	0.1688	27.2345
	DIIVINE	0.7038	0.5724	0.6958	25.5244
	NIQE	**0.9372**	*0.7258*	*0.9357*	*11.5852*
	Choi *et al.* [164]	0.6900	0.4679	0.6867	31.9258
	Balboa *et al.* [165]	0.3958	0.2486	0.4138	35.2584
NR-IQA Methods for Underwater Images	UCIQE	*0.9005*	0.6854	0.8992	16.1584
	PUIQ	0.5968	0.3589	0.6002	33.4852
	MSLBP-UWIQA	0.8286	**0.9502**	**0.9475**	**8.4735**

4.2 NR-UWIQA METRIC BASED ON A CNN ARCHITECTURE THAT USES SALIENT PATCHES

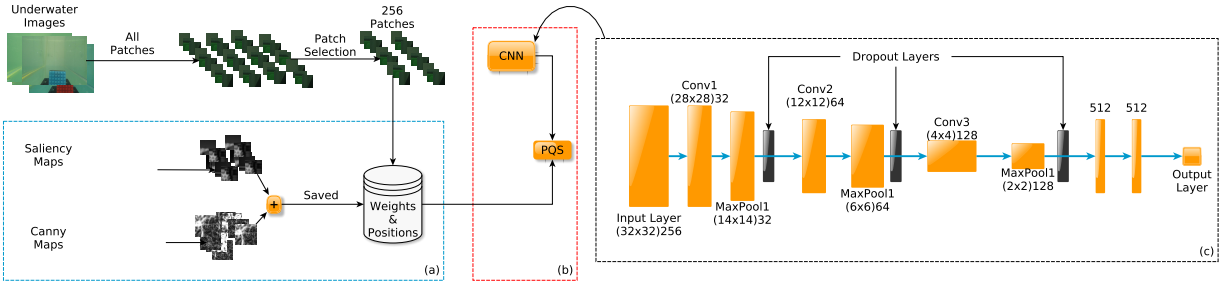

Figure 4.5: Block diagram of the MSLBP NR UWIQA method: (a) the process of selecting the most perceptually relevant patches, (b) the process of computing the predicted quality score, (c) the adapted version of VSBIQA model.

In recent years, deep learning has achieved very good results in computer vision-related tasks such as object detection, image recognition, and image segmentation. Many researchers applied deep learning to underwater vision to solve the problem related to underwater image processing. We have designed a deep learning NR-UWIQA metric, using a deep convolutional neural network (CNN) architecture that takes into consideration the visual saliency of an underwater image. Figure 4.5 shows the block diagram of the proposed NR UW-IQA method. We use the VSBIQA CNN architecture proposed by Li and Zou [101]. We chose this architecture for its good performance in similar applications. The VSBIQA has a total of nine layers. The input layer takes as input 32×32 patches in RGB color format. The second, fourth, and sixth layers are convolution layers, with stride sizes of 5×5, 3×3, and 3×3, respectively. The third, fifth, and seventh layers are MaxPool layers of size 2×2. All layers are activated by a Rectified Linear Units (ReLU) activation function. Despite its name, it is not linear and provides better performance by replacing

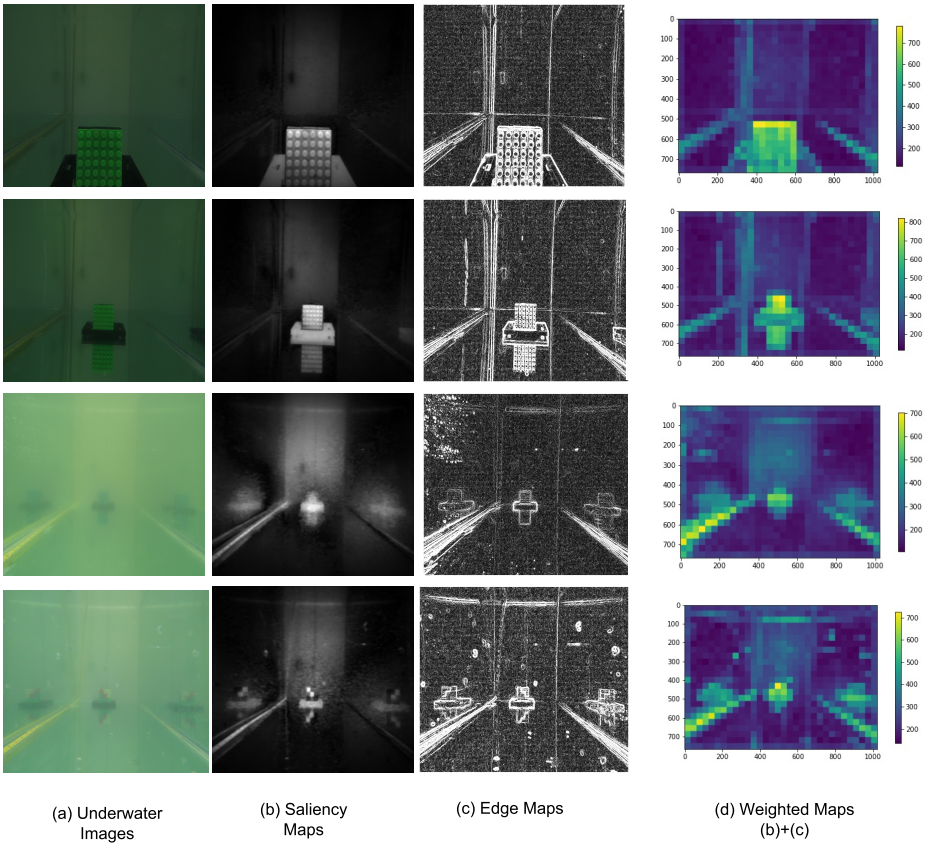

Figure 4.6: (a) Examples of underwater images taken from the UID-LEIA dataset; (b) Saliency maps of (a); (c) Edge maps of (a); and (d) Weighted maps of (b) and (c).

negative values with zeros. As shown in Figure 4.5(c), we have adapted the VSBIQA model so that every MaxPool layer is followed by a Dropout layer with a rate value of 0.5. Due to the limited size of the underwater image datasets, dropout layers are added to help reduce overfitting.

As shown in Figure 4.5(a), instead of processing a complete underwater image, the proposed method selects the most perceptually relevant regions of each underwater image and crops that region into small patches. The selected patches are then forwarded to CNN. To select the patches that are most perceptually relevant, we computed (1) the salient areas of the image and (2) the edges of the image. To compute the salient areas of the image, we use the SDSP bottom-up visual attention model [166], which has low computational cost and good performance. The SDSP algorithm has three main steps. First, it extracts features from the picture frames using a band-pass filter. The frames are then converted to CIE $L^*a^*b^*$ and filtered using a log-Gabor filter. Finally, all extracted features are combined to compute a saliency map. Figure 4.6(b) shows the SDSP saliency maps of the underwater images in Figure 4.6(a). To extract the edges of the image, we use the Canny algorithm [167] [168]. Figure 4.6(c) shows the Canny edge maps of the underwater images in Figure 4.6(a).

Let $SS(i, j)$ be the value of the saliency map at position (i, j), while $TS(i, j)$ be the value of

the edge map at the same position. We subdivide the saliency and edge maps into patches P_ℓ of size 32×32, as shown in Figure 4.5(a). The amount of saliency and edge information in the ℓ-th patch is given by:

$$\begin{aligned} SS_{P_\ell} &= \sum_{(i,j)\in P_\ell} SS(i,j), \\ TS_{P_\ell} &= \sum_{(i,j)\in P_\ell} TS(i,j). \end{aligned} \tag{4.7}$$

The relevance weight of the ℓ-th patch is defined as [101]:

$$W_{P_\ell} = \alpha \cdot SS_{P_\ell} + \beta \cdot TS_{P_\ell}, \tag{4.8}$$

where α and β are constant values that balance the contributions of saliency and edge information. In this work, we tested several values for these constants and found that $\alpha = 0.4$ and $\beta = 0.6$ [101] provide a good representation between these two types of information. Figure 4.6(d) shows the heatmaps images of the combined saliency and edge information of the underwater images in Figure 4.6(a), where brighter colors correspond to more important areas and therefore acquire higher weights. Heat maps are two-dimensional graphical representations of weights, where each value of the matrix is displayed as a color. Higher weights are represented by lighter colors, where lower weights are represented by dark colors, as shown in right-side vertical bar of Figure 4.6(d).

The patch selection process consists, first, of sorting all patches in decreasing order of W_{P_ℓ} and then choosing the most relevant patches L. In this work, we used the highest weighted 256 patches ($L = 256$) [169] of each underwater image, which is a good compromise in terms of complexity and information representation [101]. As shown in Figure 4.5(b), the predicted quality score corresponding to each underwater image (PQS_{UI}) is obtained by computing a weighted average of the predicted quality scores for each patch [101], as given by the following equation:

$$PQS_{UI} = \frac{\sum_{\ell=1}^{L} W_{P_\ell} \cdot PQS_{P_\ell}}{\sum_{\ell=1}^{L} W_{P_\ell}}, \tag{4.9}$$

where L is the number of selected patches, W_{P_ℓ} is the weight of the ℓ-th patch (P_ℓ), and PQS_{P_ℓ} is the predicted quality score of P_ℓ.

We compare the proposed methods with the following underwater IQA methods: (i) the Underwater Color Image Quality Evaluation (UCIQE) [56] and (ii) the Patch-Based Mean Underwater Image Quality (PUIQ) [57]. Furthermore, we also compare our method with traditional NR-IQA methods: (a) CORNIA [116], (b) BRIQUE [119], (c) SSEQ [118], (d) DIIVINE [121], (d) NIQE [163], Choi *et al.* [164], and (e) Balboa *et al.* [165]. The experiments were carried out on a PC with an Intel Core i7-4790 processor at 3.60GHz, running an Ubuntu operating system.

As mentioned before, in this work, we used the UID-LEIA dataset, which contains 135 distorted underwater images. Although this is a small number of input samples for training a CNN, this is the only data set that contains subjective quality scores (mean opinion scores - MOS) associa-

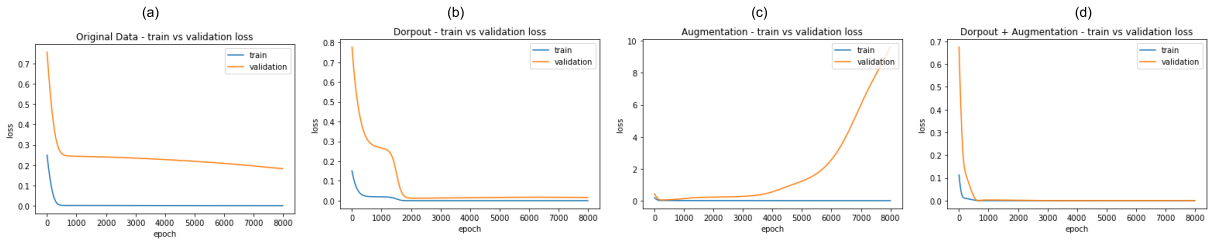

Figure 4.7: Training vs Validation Loss curves: (a) CNN performance on original data without adding dropout layers, (b) CNN performance on original data after adding Dropout layers, (c) CNN performance without Dropout on augmented data, and (d) CNN performance after adding Dropout layers on augmented data.

ted with each image. Therefore, this is one of the only datasets that can be used in the design of an underwater quality metric. To increase the size of the training data, we augment the input samples to achieve good performance without overfitting problems. More specifically, we used horizontal flip, vertical flip, and four different shifts at the top-left, top-right, bottom-left, and bottom-right corners of the images. In this way, we obtained 810 distorted underwater images.

To train and test CNN, we used the K-fold cross-validation approach that splits the dataset into K consecutive folds. In our simulations, we set $K = 10$ and perform 10-fold simulations. In each simulation, one fold is used as a test set, while the remaining $K - 1 = 9$ folds are used as the training set. After the 10-fold simulations, we report the mean correlation values. As mentioned earlier, we selected at most 256 patches ($L = 256$) from each underwater image.

Figure 4.7 shows the performance loss results of the proposed method for different test scenarios. Figure 4.7(a) shows the training and validation loss curves obtained for the original CNN architecture, without the dropout layers. The graph indicates an overfitting, since there is a large difference between the train and validation loss curves. Figure 4.7(b) shows the loss curves when we add dropout layers to the CNN architecture. Notice that the curves still show an overfitting effect. Figure 4.7(c) shows the loss curves for the original CNN architecture (without adding Dropout layers) with data augmentation. Notice that after 1000 epochs, the graph starts showing some overfitting. Finally, Figure 4.7(d) shows the loss curves for the CNN architecture with dropout layers and using data augmentation. This graph shows very close validation and train-loss curves, indicating that there are no clear signs of overfitting.

Table 4.2 illustrates the performance comparison of the proposed method (CNN + dropout layers trained on augmented data) with other state-of-the-art metrics in the UID-LEIA dataset. Values in bold correspond to the best performance, whereas results in italics correspond to the second-best results for the SROCC and PLCC columns. The highest PLCC value is obtained by the M. Irshad *et al.* [62] metric, which is an algorithm based on ML, while the highest SROCC values are obtained by the proposed method.

Figure 4.8 shows the box plot of average SRCC and PLCC values across 10-fold simulations for all metrics. Since the authors of the UW-IQA methods did not publish their code, we compared the results of 10-fold with only general purpose NR-IQA methods. For methods that are not based

Table 4.2: Performance evaluation of proposed methods with different IQA methods on UID-LEIA dataset.

Type	Method	PLCC	SROCC
General Purpose NR-IQA Methods	CORNIA	0.5463	0.6203
	SSEQ	0.1134	0.1286
	BRISQUE	0.2703	0.3325
	DIIVINE	0.5302	0.6365
	NIQE	0.5917	0.7227
	Choi *et al.* [164]	0.4679	0.6867
	Balboa *et al.* [165]	0.2486	0.4138
NR-UW-IQA Methods	UCIQE	0.6854	0.8992
	PUIQ	0.3589	0.6002
ML and CNN Based NR-UW-IQA Methods	MSLBP-UWIQA	**0.9502**	*0.9475*
	CNN-SP-UWIQA	*0.9426*	**0.9504**

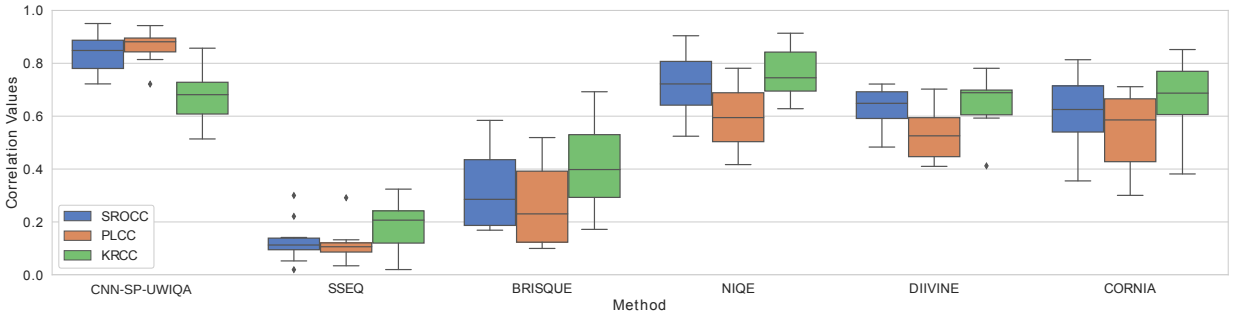

Figure 4.8: Box plot of SROCC and PLCC results obtained by CNN-SP-UWIQA method on UID-LEIA dataset.

on training, we adopted the same split strategy, but considered only the 20% test set and discarded the 80% train set. Notice that among all metrics, the SROCC and PLCC results provided by the proposed method are the highest and have the smallest variation, which means that the results of our method are very consistent. From Figure 4.8, it can be observed that the metrics SSEQ and BRISQUE have shown the worst performance. For BRISQUE and DIIVINE, the boxes have wide range area, which represents inconsistency in performance.

We analyze whether the selected patches with their corresponding weights (computed by combining the saliency and edge maps in Equation 4.9) performed better than using all patches without their weights. For this purpose, we performed an ablation test, in which we trained and tested the model using all patches of underwater images without their weights. In other words, we remove the block (a) from Figure 4.5. Table 4.3 shows the SROCC and PLCC values obtained, which are significantly lower than the results shown in Table 4.2. In other words, using a model with selected patches and their corresponding weights provides better performance than using all patches without their weights.

Table 4.4 presents the time required to train and test the proposed method, using the UID-LEIA dataset. We compare the time consumption with M. Irshad *et al.* method, which corresponds to using a simple Random Forest regression algorithm. Notice that M. Irshad *et al.* method requires 61 seconds for pre-processing (loading images in RGB format), 3.1 minutes for training, and 0.3 seconds for testing the model. On the other hand, since the proposed method is based on

Table 4.3: Comparison of CNN-SP-UWIQA model with a model processing all patches of underwater images without their weights. Training/Test is performed on the UID-LEIA dataset.

Dataset	Test Type	SROCC	PLCC
UID-LEIA	All patches without their weights	0.8309	0.7914
	Selected patches with their corresponding weights	**0.9426**	**0.9504**

Table 4.4: The time consumption of the proposed methods on UID-LEIA dataset.

Method	Pre-Processing (seconds)	Training (minutes)	Testing (seconds)
MSLBP-UWIQA	61	3.1	0.3
CNN-SP-UWIQA	60	16	0.42

a multilayered convolutional neural network, it requires a slightly longer time for training and testing, but a small amount of time is consumed for data reprocessing.

4.3 CONCLUSIONS

In this chapter, we have presented two quality metrics for underwater images. The first method used an adapted version of the multiscale salient local binary pattern operator to extract image features and a machine learning approach to predict quality. In the second method, we trained a CNN model using patches of underwater images. we chose patches that were most perceptually relevant by incorporating the properties of the human visual system. These selected patches were then fed to the CNN model for quality estimation. The results show that the proposed metrics are efficient and fast, and can be implemented in real time for different underwater image quality applications.

5 TESTING THE PROPOSED NR-UWIQA METRICS ON REAL UNDERWATER IMAGES

This chapter discusses the development of an underwater image enhancement database. Underwater image enhancement has been attracting much attention due to its importance in sea exploration. In recent years, many underwater image enhancement algorithms have been proposed to improve underwater image quality. However, how to fairly compare the performance of underwater image enhancement algorithms, comparing underwater image enhancement algorithms remains a difficult problem. The lack of a comprehensive subjective user study with a large benchmark dataset and a reliable objective image quality assessment (IQA) metric has made it difficult to fully understand the actual performance of underwater image enhancement algorithms. In this work, we aim to fill these gaps in both subjective and objective terms. First, we construct a new subjectively annotated underwater image enhancement benchmark dataset (UIEB) that simultaneously contains raw real underwater images, readily available enhancement results from representative underwater image enhancement algorithms, and subjective evaluation results for each enhanced underwater image. In the second part, we present an effective NR Underwater Image Quality metric for automatically assessing the quality of enhanced underwater images.

5.1 GENERATION OF REAL UNDERWATER IMAGES

Underwater images are inevitably affected by wavelength-dependent absorption and scattering. Consequently, raw underwater images usually suffer from various quality deficiencies such as color distortion, uneven illumination, reduced contrast and visibility, etc., which significantly affect the performance of underwater application systems [170]. To alleviate these problems, numerous underwater image enhancement (UIE) algorithms have been proposed in the literature. However, a difficult and unsolved problem is how to fairly compare the performance of different UIE algorithms or evaluate the quality of the enhanced results. We believe that the performance comparison of existing work has obvious limitations. First, the qualitative comparison is not entirely convincing because only a limited number of samples are presented and, more importantly, researchers may select different samples for visual comparison. Second, the accuracy of the existing quality metrics used for the quantitative comparison is not high. As shown in the experiments, the existing objective quality metrics do not fully match the human subjective perception. Therefore, a quantitative comparison using these metrics cannot accurately reflect the performance of underwater image enhancement algorithms. Clearly, there is a need for a large-scale underwater image enhancement dataset that can be used for a fair and comprehensive performance evaluation of underwater image enhancement algorithms.

In the past few years, the use of machine learning and deep learning techniques to design

image quality assessment methods has become extremely popular. To design a metric, we need a ground-truth dataset to assess the quality of underwater images. In this work, we propose a new database for the assessment of underwater image quality to advance research in this area. We have taken enhanced real underwater images from the Underwater Image Enhancement Benchmark (UIEB) dataset [171] have diverse color ranges and degree of contrast. The corresponding reference images are of relatively genuine color and have improved visibility and brightness. Figure 5.1 shows the reference underwater images. There are 200 real underwater images processed with different enhancement algorithms in this collection. We present the perception study and design a subjective online experiment for the mean opinion score (MOS) of reference underwater images. The results from MOS can be used to effectively test underwater image enhancement algorithms and to develop underwater IQA metrics. This dataset can be freely downloaded and used for scientific research.

Figure 5.1: Example of reference underwater images from UIEB dataset

5.1.1 A Crowdsourcing Experiment for Underwater Image Quality Assessment

To perform a crowdsourcing experiment, we have selected a Single Stimulus (SS) approach in which the test images are randomly displayed. Then, observers are asked to rate the quality of the displayed image using a category rating scale. This quality scale is known as the Absolute Category Rating Scale (ACR). We conducted this subjective experiment online using the QuickEval web platform [172]. We have chosen 200 real underwater images from the UIEB database, and these images are enhanced by different enhancement algorithms. A total of 12 image enhancement techniques are employed to generate the potential reference images, including nine underwater techniques that are fusion based [45], two-step-based [173], retinex-based [174], UDCP [175], regression-based [176], GDCP [177], Red Channel [178], histogram prior [179] and blurriness-based [51] two image dehazing methods that are DCP [180] and MSCNN [181] and one commercial application for enhancing the underwater image is dive+8.

We chose to do the online experiment, although there is a debate on the conditions of the online versus the offline experiment. The offline experiments are controlled, which is good for reproducibility purposes, while there can be variations in screen resolution, lighting, and viewing conditions during online experiments. Performing an online experiment allows for more and more diverse observers, which is important for statistical significance of the data. For each test stimulus, the SS experiment outputs a mean opinion score (MOS) value, which is computed by taking the average of the ratings for all observers. Figure 5.2 shows the interface of the QuickEval online system.

A total of 40 participants participated in the online subjective experiment, with 62% of the participants being male and 38% being female. The average age of the participants was 32 for males and 34 for females. To analyze the subjective data gathered, we computed the mean opinion score (MOS).

5.1.2 Testing NR-IQA Metrics on the UIEB Dataset

To analyze the performance of general-purpose image quality metrics with respect to the MOS obtained in subjective experiment for the UIEB dataset, we performed quantitative and qualitative comparisons. For quantitative comparison, the performance comparison results of the original objective quality metrics in the UIEB dataset are summarized in Table 5.1. The results present the correlation with respect to SROCC and PLCC between the MOS scores and the original objective quality metrics, and are computed by the basic image quality metrics SSIM [182], PSNR, DIIVINE [121] and BRISQUE [119]. The results show that the DIIVINE method provides the best performance in terms of SROCC (0.5138) and PLCC (0.4865), compared to the BRIQSUE, SSIM, and PSNR methods. Both the BRISQUE and DIIVINE methods employ natural scene statistics (NSS) from images, but BRISQUE works in the spatial domain, while DIIVINE works in the frequency domain.

For a qualitative comparison, Figure 5.3 shows the scatter plots of all metrics to visually il-

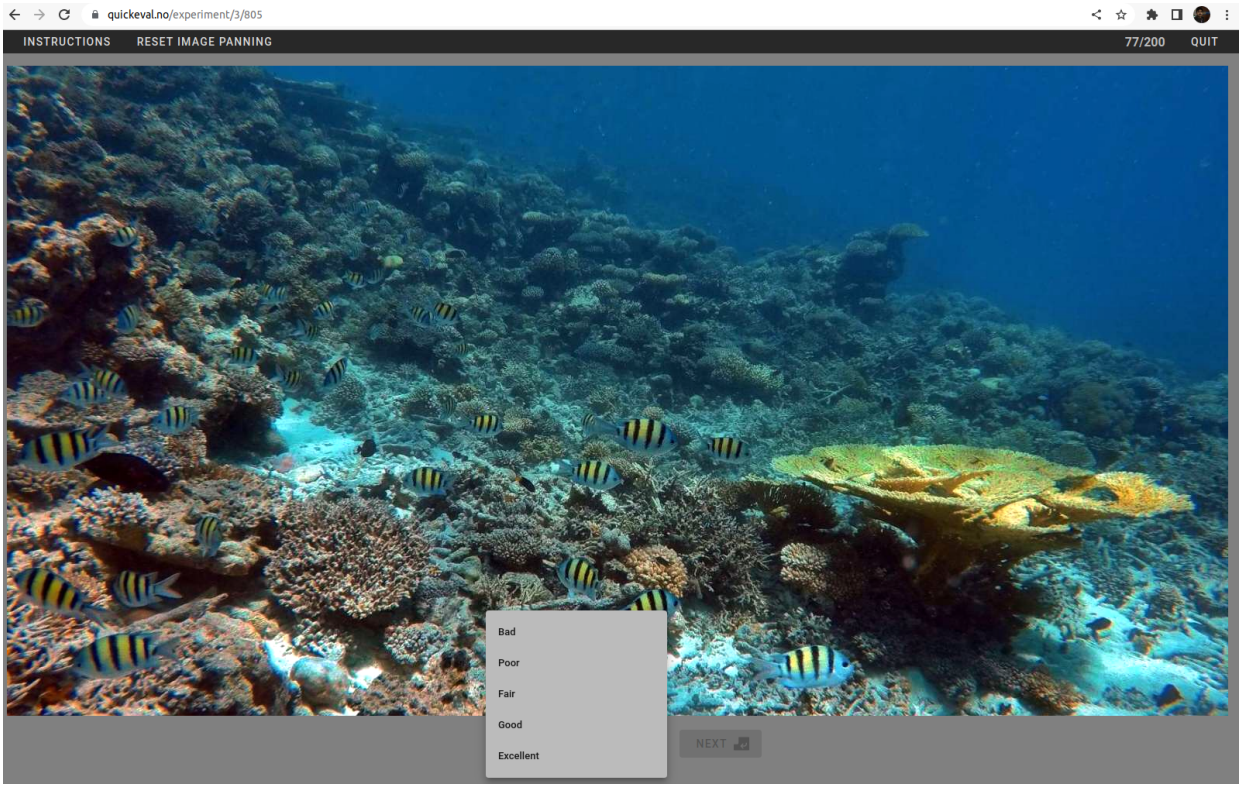

Figure 5.2: Interface of QuickEval subjective experiment system

Table 5.1: Objective quality evaluation of the UIEB dataset in comparison with image quality metrics.

Type	Method	PLCC	SROCC
IQA Methods	BRISQUE	0.1733	0.2071
	PSNR	0.2162	0.1927
	SSIM	0.4044	0.4088
	DIIVINE	**0.5138**	**0.4865**

lustrate the prediction ability of the objective quality metrics evaluated. The plots show similar results, as described in Table 5.1, but all objective quality metrics individually are not good predictors for the UIEB dataset, as indicated by the wide spread of metric values in the fitted lines. To elaborate, the blue asterisks in the scatter plots represent a paired measurement of two variables (objective quality metric and MOS), while the red lines are the line of best fit of each objective quality metric. If the distance between the blue asterisks and the red lines is far away or nonlinear, it means that the variables are weakly correlated or not correlated. If the variables are strongly correlated, the blue asterisks should cluster close to the red line. In other words, the predicted performance of a quality metric with a tight scatter of data about the line of best fit will be superior to that of a quality metric with a wider scatter.

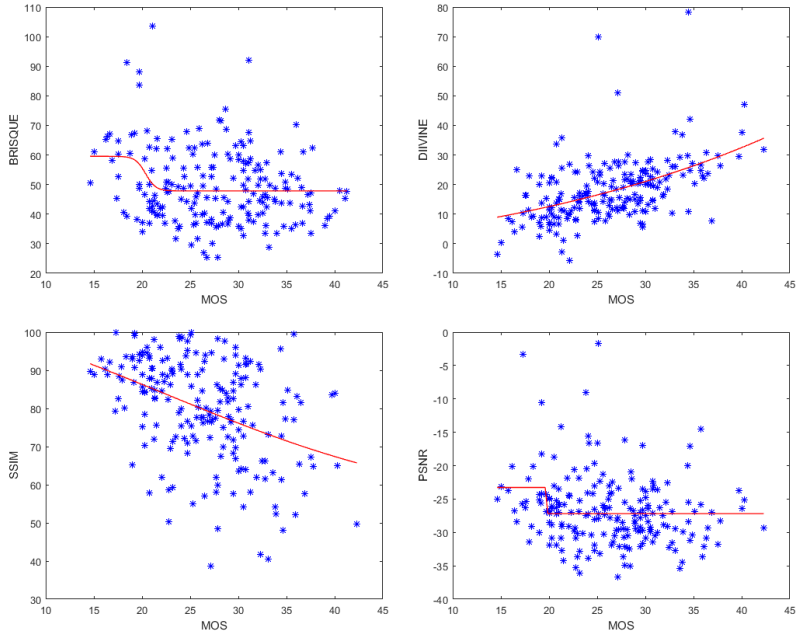

Figure 5.3: Scatter plots of all compared objective metrics on the UIEB dataset. The blue asterisks in the scatter plots represent a paired measurement of two variables (Objective quality metric and MOS), while the red lines are the fitted linear function on the UIEB dataset of each objective quality.

5.1.3 Testing the Proposed NR-UWIQA Metrics on the UIEB Dataset

In this section, we tested the two proposed metrics (MSLBP-UWIQA and CNN-SP-UWIQA) on the UIEB datase. To increase the size of training data and achieve good performance without overfitting problems, in this work we have used data augmentation techniques to increase the number of input samples. More specifically, we used horizontal and vertical flips. In this way, a total of 400 distorted underwater images are obtained. Another measure we took to avoid overfitting is adding a Dropout layer to the CNN architecture.

For illustration, the performance of the CNN-SP-UWIQA metric, in Figure 5.4(b) shows the SDSP saliency maps of underwater images in Figure 5.4(a), while in Figure 5.4(c) shows the corresponding Canny edge maps of these underwater images. Figure 5.4(d) shows the heatmaps images of the combined saliency and edge information of the underwater images in Figure 5.4(a), where brighter colors correspond to more important areas and therefore acquire higher weights.

To train and test the CNN of the CNN-SP-UWIQA, we used the K-fold cross-validation approach that splits the dataset into K consecutive folds. In our simulations, we set $K = 10$ and perform 10-fold simulations. In each simulation, one fold is used as a test set, while the remaining $K - 1 = 9$ folds are used as the training set. After the 10-fold simulations, we report the mean correlation values. As mentioned above, we selected at most 256 patches ($L = 256$) from each underwater image. For training the model, we used the following parameters: mini-batches of size 1, and mean squared error (MSE) as the training loss. Furthermore, the Adam optimizer [183] is used to minimize the loss function. We initially set epochs to 6,000, and monitor the training

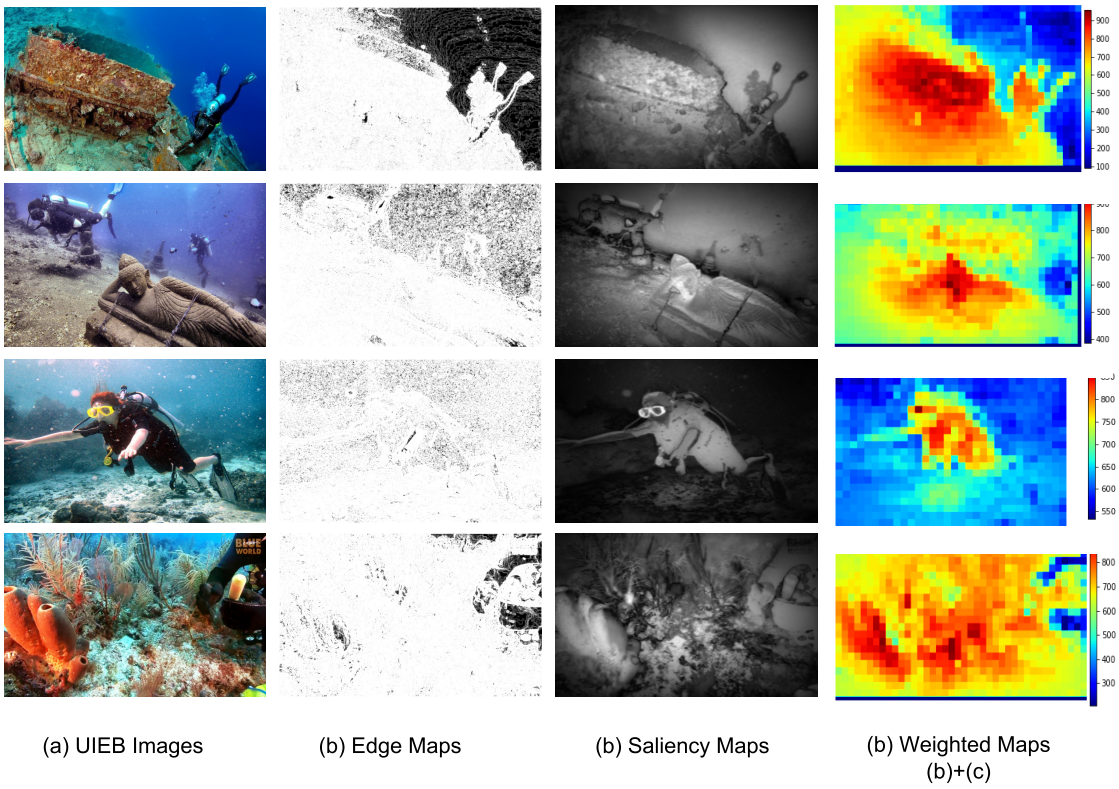

Figure 5.4: (a) Examples of underwater images taken from the UIEB dataset; (b) Edge maps of (a); (c) Saliency maps of (a); and (d) Weighted maps of (b) and (c).

Table 5.2: Performance evaluation of proposed methods with different IQA methods on UIEB dataset.

Type	**Method**	**PLCC**	**SROCC**
General Purpose IQA Methods	SSIM	0.4546	0.4815
	BRISQUE	0.1263	0.1511
	PSNR	0.2766	0.2413
	DIIVINE	0.4860	0.4785
Proposed Methods	CNN-SP-UWIQA	**0.9619**	**0.9356**
	MSLBP-UWIQA	0.6002	0.6618

loss by applying the early stopping method with 200 patience. We considered a dropout layer and augmented data.

For the MSLBP-UWIQA method, we adopted the k-fold cross-validation strategy in which $k = 1000$, and from the UIEB dataset, 80% of the content is used for training and 20% for testing. This 80-20 training-testing random split is performed 1,000 times, and then the mean correlation is computed and reported the average correlations. The experiments were carried out on a PC with an Intel Core i7-4790 processor at 3.60GHz with 32 GB GPU, running an Ubuntu operating system. All experiments are implemented using the Chainer framework in Python.

We compare the NR-UWIQA methods with the traditional FR and NR IQA methods: (a) BRISQUE [119], (b) DIIVINE [121], (c) SSIM [182], (d) PSNR. Table 5.2 illustrates the perfor-

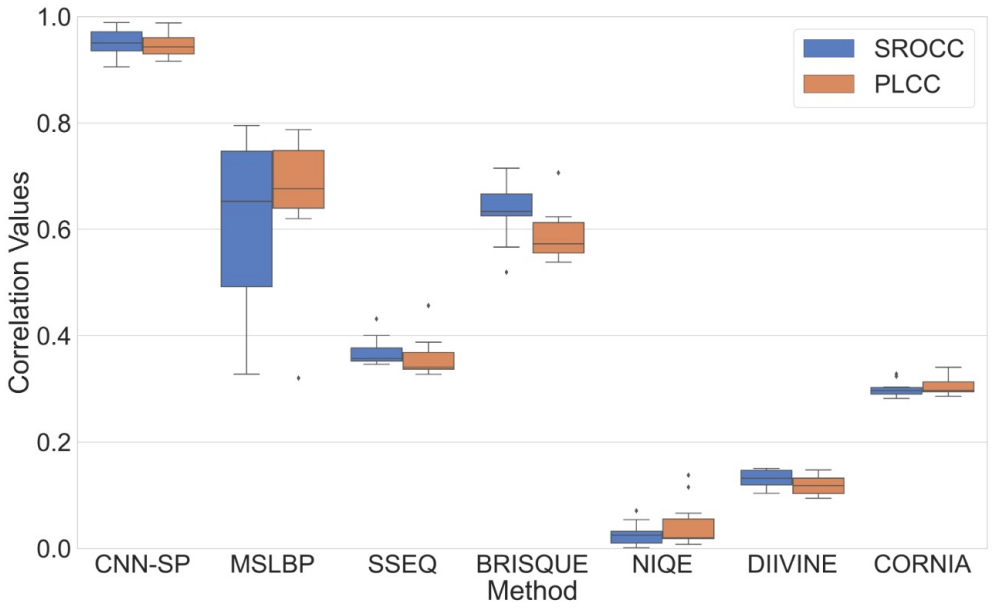

Figure 5.5: Box plot of SROCC and PLCC results obtianed by CNN-SP-UWIQA method on UIEB dataset.

mance comparison of the proposed methods with the comparison metrics for the UIEB dataset. We report the average correlations of 10-folds simulations using all comparison metrics on the UIEB dataset. Values in bold correspond to the best performance with respect to the SROCC and PLCC columns. According to the results, the proposed methods achieved the highest correlation values on real underwater images.

Figure 5.5 shows the box plot of the average values of SROCC and PLCC in the 10-fold simulations for all metrics. Note that among all metrics, the SROCC and PLCC UIEDresults provided by the the CNN-SP-UWIQA metric are the highest and have the smallest variation, which means that the results of our method are very consistent. In Figure 5.5, it can be seen that the PSNR and BRISQUE metrics have shown the worst performance. Notice that the MSLBP-UWIQA method, the SROCC and PLCC boxes have a wide range area, which represents an inconsistency in performance.

Next, we take a closer look at the performance of the CNN-SP-UWIQA metric. First, we analyze whether feeding the metric with only a selection of the most salient patches (computed by combining the saliency and edge maps in Equation 4.8) results in better performance than feeding all patches. For this purpose, we performed an ablation test, in which we trained and tested the model using all patches of underwater images. In other words, we removed the block (a) shown in Figure 4.5. We also performed a test in which we trained and tested the model after removing the Dropout layer.

Table 5.3 shows the SROCC and PLCC values obtained for these tests, which shows that the correlation values obtained using all patches are significantly lower than the results obtained by selecting only the most salient patches. As mentioned earlier, the purpose of using saliency-based information is to eliminate any redundancy, keeping only the information that the human eye considers more relevant. Weighting the most salient regions (and therefore distortions) is a

Table 5.3: Comparison of the proposed CNN-SP-UWIQA with a model of different variants. Training/Test is performed on the UIEB dataset.

Dataset	**Test**	**SROCC**	**PLCC**
UIEB	All patches (no selection)	0.7099	0.6861
	Without Dropout layer	0.4822	0.3250
	Selected most salient patches	**0.9619**	**0.9356**

Table 5.4: The time consumption of the proposed CNN-SP-UWIQA metric on UIEB dataset.

Method	**Pre-Processing** (seconds)	**Training** (minutes)	**Testing** (seconds)
MSLBP-UWIQA	92	0.22	0.05
CNN-SP-UWIQA	42	113	7.5

popular approach in image quality assessment, with several saliency-based image quality metrics proposed over the past decade [184, 185]. Therefore, this result shows that a similar approach can be used with CNN architectures for UWIQA methods. In other words, by creating a CNN model that takes as input only selected salient patches, we provide the network with the most perceptually relevant visual information, filtering out unimportant areas, and obtaining a better accuracy performance. In Table 5.3, we also see that the model without the dropout layer diverges and performance decreases. By adding the dropout layer, the model provides the best fit for the test dataset.

Table 5.4 presents the running time required to train and test the NR-UWIQA method, using the UIEB dataset. We compare the time consumption of CNN-SP-UWIQA with the time consumption of the MSLBP-UWIQA method, which uses a simple Random Forest regression algorithm. Note that the CNN-SP-UWIQA method takes less time to prepare input features. More specifically, the MSLBP-UWIQA method requires 92 seconds, while the CNN-SP-UWIQA method requires 42 seconds for preprocessing. For training and testing, since the CNN method incorporates a multilayered convolutional neural network, its computational cost is higher compared to the MSLBP-UWIQA method.

5.2 CONCLUSIONS

In this chapter, We have developed a UIEB database that is used for evaluating the performance of underwater image quality assessment methods. To collect the subjective quality scores of UIEB dataset, We performed a crowdsourcing experiment in which the observers were asked to rate the quality of the displayed image using a category rating scale. We conducted this subjective experiment online using the QuickEval web platform. We analyzed the performance of general-purpose image quality metrics with respect to the MOS obtained in subjective experiment for the

UIEB dataset. We also performed quantitative and qualitative comparisons.

We tested the proposed metrics (MSLBP-UWIQA and CNN-SP-UWIQA) on the UIEB database. To train the CNN model we used the K-fold cross validation approach that splits the dataset into K consecutive folds. In training, instead of feeding the model with patches selected randomly from underwater images, we chose patches that were most perceptually relevant by incorporating the properties of human visual system (HVS) name as edge and visual saliency. For the MSLBP-UWIQA method, we adopted the k-fold cross-validation strategy in which $k = 1000$, and from the UIEB dataset, 80% of the content is used for training and 20% for testing. The Performance evaluations of proposed methods are good compared to general purpose IQA methods. Results suggest the metric can be used in real-time to assess underwater image quality for a number of applications, and that it is efficient and fast.

6 CONCLUSIONS AND FUTURE WORK

6.1 CONCLUSIONS

In this work, our goal was to investigate the quality assessment of the enhanced images. Image enhancement algorithms have the goal of improving the image quality and, therefore, the usefulness of an image for a given task. Although there are several image enhancement algorithms, there is no consensus on how to estimate the performance of these enhancement algorithms. Since the final consumers of the resulting enhanced visual content are human viewers, the performance of these algorithms should take into account the perceived visual quality of the resulting enhanced images. Unfortunately, although in the last decades a lot of progress has been made in the area of image quality assessment, designing metrics to estimate the quality of enhanced and restored images remains a challenge. This is particularly true for underwater image application, where images frequently need to be restored because of the severity of the degradations introduced by the underwater environment. Therefore, there is a great need for quality metrics that can estimate the quality of enhanced and restored images. In this thesis, our goal is to design metrics for this scenario.

First, we presented a psychophysical / subjective study in which we analyzed the perceptual quality of images that were enhanced with several types of enhancement algorithms, including color, sharpness, histogram, and contrast. Specifically, we enhanced 35 source images obtained from publicly available databases. Then, we performed a subjective experiment to assess the quality of the enhanced images. We used a Double Stimulus Continuous Quality Scale (DSCQS) experimental methodology, with a between-subject approach where each subject scored a subset of the total database to avoid fatigue. Given the large number of test images, we designed a crowd-sourcing interface to perform an online subjective experiment. This type of interface has the advantage of making it possible to collect data from many participants. We also performed an experiment in a controlled laboratory environment and compared its results with the crowdsourcing results. Since there are very few quality enhancement databases available in the literature, this work represents a contribution to the area of enhanced image quality.

Second, we have presented a lightweight no-reference (NR) underwater image quality assessment evaluation metric for enhanced underwater images (NR-UWIQA), which is based on a machine learning approach to predict quality and an adapted version of the multiscale salient local binary pattern operator. The proposed metric was tested on the UID-LEIA database and presented good accuracy performance when compared to other state-of-the-art methods. In summary, the proposed NR-UWQIA method can be used to assess the results of restoration techniques quickly and efficiently, opening up a new perspective in the area of underwater image restoration and quality assessment.

Moreover, we present a lightweight no-reference (NR) underwater image quality assessment

method based on convolutional neural networks (CNNs). We trained the CNN model using patches of underwater images. Instead of feeding the model with patches selected randomly from underwater images, we chose patches that were most perceptually relevant by incorporating the properties of human visual system (HVS) name as edge and visual saliency. These HVS properties helped to calculate the regions of interest that were used as training data. The proposed method has greater efficiency and robustness. Experimental results on the underwater image database demonstrate that our approach outperforms the state-of-the-art methods compared.

Furthermore, we developed a new subjectively annotated underwater image enhancement benchmark dataset (UIEB) that simultaneously contains raw real underwater images, readily available enhancement results from representative underwater image enhancement algorithms. For subjective evaluation we perform a crowdsourcing experiment, we have selected a Single Stimulus (SS) approach in which the test images are randomly displayed. Then, observers are asked to rate the quality of the displayed image using a category rating scale. This quality scale is known as the Absolute Category Rating Scale (ACR). We conducted this subjective experiment online using the QuickEval web platform. In the second part, we present an effective NR underwater image quality metric for automatically assessing the quality of enhanced underwater images. We also analyzed the performance of general-purpose image quality metrics with respect to the MOS obtained in subjective experiment for the UIEB dataset. We also performed quantitative and qualitative comparisons.

We tested the proposed metrics (MSLBP-UWIQA and CNN-SP-UWIQA) on the UIEB database. To train the CNN model we used the K-fold cross validation approach that splits the dataset into K consecutive folds. In training, instead of feeding the model with patches selected randomly from underwater images, we chose patches that were most perceptually relevant by incorporating the properties of human visual system (HVS) name as edge and visual saliency. For the MSLBP-UWIQA method, we adopted the k-fold cross-validation strategy in which $k = 1000$, and from the UIEB dataset, 80% of the content is used for training and 20% for testing. The Performance evaluations of proposed methods are good compared to general purpose IQA methods. Results suggest the metric can be used in real-time to assess underwater image quality for a number of applications, and that it is efficient and fast.

6.2 FUTURE WORK

In the first subjective experiment, we obtained a dataset of enhanced images with the corresponding DMOS. As part of our future work, we intend to perform a statistical analysis using objective quality metrics. Also, we are working on designing no-reference metrics specifically to assess the quality of enhanced images.

Another possible future work is to propose a new deep neural network-based quality assessment metric for underwater and enhanced images. In this work, we only consider convolutional

neural networks. However, in real-world scenarios, different types of deep learning methods such as atrous convolution layers, long- and short-term memory networks have produced good results in predictions [186, 187, 188]. Therefore, it would be interesting to determine what features of enhanced and underwater images can be used to detect and estimate the strength of these deep learning methods.

Given the current benchmark of underwater enhanced images, there is a limitation of access. To remove this obstacle, our future work includes to establish a dataset of underwater enhanced images. Then, a subjective methodology will be carried out to obtain subjective quality scores for every image. The dataset will serve as a motivation towards the development of more quality metrics for this content media.

BIBLIOGRAPHIC REFERENCES

1 SÁNCHEZ-FERREIRA, C.; COELHO, L.; AYALA, H. V.; FARIAS, M. C.; LLANOS, C. H. Bio-inspired optimization algorithms for real underwater image restoration. *Signal Processing: Image Communication*, Elsevier, v. 77, p. 49–65, 2019.

2 FREITAS, P. G.; AKAMINE, W. Y.; FARIAS, M. C. Blind image quality assessment using multiscale local binary patterns. *Journal of Imaging Science and Technology*, Society for Imaging Science and Technology, v. 60, n. 6, p. 60405–1, 2016.

3 Ghadiyaram, D.; Bovik, A. C. Massive online crowd sourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, v. 25, n. 1, p. 372–387, Jan 2016. ISSN 1941-0042.

4 PONOMARENKO, N.; JIN, L.; IEREMEIEV, O.; LUKIN, V.; EGIAZARIAN, K.; ASTOLA, J.; VOZEL, B.; CHEHDI, K.; CARLI, M.; BATTISTI, F.; KUO, C.-C. J. Image database tid2013: Peculiarities, results and perspectives. *Signal Processing: Image Communication*, v. 30, p. 57 – 77, 2015. ISSN 0923-5965. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0923596514001490>.

5 SEGHIR, Z. A.; HACHOUF, F. Full-reference image quality assessment measure based on color distortion. In: AMINE, A.; BELLATRECHE, L.; ELBERRICHI, Z.; NEUHOLD, E. J.; WREMBEL, R. (Ed.). *Computer Science and Its Applications*. Cham: Springer International Publishing, 2015. p. 66–77. ISBN 978-3-319-19578-0.

6 CHANDLER, D. M. Seven challenges in image quality assessment: past, present, and future research. *ISRN Signal Processing*, Hindawi Publishing Corporation, v. 2013, 2013.

7 CISCO. *Cisco Visual Networking Index: Global Mobile Data Traffic Forecast Update, 2017–2022*. 2019.

8 ITU-R:BT.500-13. *methodology for the subjective assessment of the quality of television pictures*. 2019. Disponível em: <https://www.itu.int/rec/R-REC-BT.500>.

9 REC, I. P. . *mean opinion score (mos) terminology,*. 2019. Disponível em: <https://www.itu.int/rec/R-REC-BT.500>.

10 SHAHID, M.; ROSSHOLM, A.; LÖVSTRÖM, B.; ZEPERNICK, H.-J. No-reference image and video quality assessment: a classification and review of recent approaches. *EURASIP Journal on image and Video Processing*, Springer, v. 2014, n. 1, p. 1–32, 2014.

11 WANG, Z.; BOVIK, A. C. Modern image quality assessment. *Synthesis Lectures on Image, Video, and Multimedia Processing*, Morgan & Claypool Publishers, v. 2, n. 1, p. 1–156, 2006.

12 Watson, A. B.; Yang, G. Y.; Solomon, J. A.; Villasenor, J. Visibility of wavelet quantization noise. *IEEE Transactions on Image Processing*, v. 6, n. 8, p. 1164–1175, Aug 1997. ISSN 1941-0042.

13 DALY, S. J. Visible differences predictor: an algorithm for the assessment of image fidelity. In: ROGOWITZ, B. E. (Ed.). *Human Vision, Visual Processing, and Digital Display III*. SPIE, 1992. v. 1666, p. 2 – 15. Disponível em: <https://doi.org/10.1117/12.135952>.

14 LUBIN, J. Digital images and human vision. In: WATSON, A. B. (Ed.). Cambridge, MA, USA: MIT Press, 1993. cap. The Use of Psychophysical Data and Models in the Analysis of Display System Performance, p. 163–178. ISBN 0-262-23171-9. Disponível em: <http://dl.acm.org/citation.cfm?id=197765.197782>.

15 WATSON, A. Visual optimization of dct quantization matrices for individual images. *Proc. AIAA Computing in Aerospace*, v. 9, 02 1993.

16 Mannos, J.; Sakrison, D. The effects of a visual fidelity criterion of the encoding of images. *IEEE Transactions on Information Theory*, v. 20, n. 4, p. 525–536, July 1974. ISSN 1557-9654.

17 You, J.; Korhonen, J.; Perkis, A. Attention modeling for video quality assessment: Balancing global quality and local quality. In: *2010 IEEE International Conference on Multimedia and Expo*. [S.l.: s.n.], 2010. p. 914–919.

18 Xin Feng; Tao Liu; Yang, D.; Yao Wang. Saliency based objective quality assessment of decoded video affected by packet losses. In: *2008 15th IEEE International Conference on Image Processing*. [S.l.: s.n.], 2008. p. 2560–2563.

19 Pessoa, A.; Falcão, A.; Nishihara, R.; Silva, A.; Lotufo, R. Video quality assessment using objective parameters based on image segmentation. *SMPTE Journal*, v. 108, n. 12, p. 865–872, Dec 1999. ISSN 0036-1682.

20 WOLF, S.; PINSON, M. H. Spatial-temporal distortion metric for in-service quality monitoring of any digital video system. In: TESCHER, A. G.; VASUDEV, B.; JR., V. M. B.; DERRYBERRY, B. (Ed.). *Multimedia Systems and Applications II*. SPIE, 1999. v. 3845, p. 266 – 277. Disponível em: <https://doi.org/10.1117/12.371210>.

21 WINKLER, S. Issues in vision modeling for perceptual video quality assessment. *Signal Processing*, v. 78, n. 2, p. 231 – 252, 1999. ISSN 0165-1684. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0165168499000626>.

22 Algazi, V. R.; Hiwasa, N. Perceptual criteria and design alternatives for low bit rate video coding. In: *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*. [S.l.: s.n.], 1993. p. 831–835 vol.2. ISSN 1058-6393.

23 WEBSTER, A. A.; JONES, C. T.; PINSON, M. H.; VORAN, S. D.; WOLF, S. Objective video quality assessment system based on human perception. In: ALLEBACH, J. P.; ROGOWITZ, B. E. (Ed.). *Human Vision, Visual Processing, and Digital Display IV*. SPIE, 1993. v. 1913, p. 15 – 26. Disponível em: <https://doi.org/10.1117/12.152700>.

24 Battisti, F.; Carli, M.; Liu, Y.; Neri, A.; Paudyal, P. Distortion-based no-reference quality metric for video transmission over ip. In: *2015 International Symposium on Signals, Circuits and Systems (ISSCS)*. [S.l.: s.n.], 2015. p. 1–4. ISSN null.

25 ZHU, K.; LI, C.; ASARI, V.; SAUPE, D. No-reference video quality assessment based on artifact measurement and statistical analysis. *Circuits and Systems for Video Technology, IEEE Transactions on*, v. 25, p. 533–546, 04 2015.

26 JIA, L.; ZHONG, X.; TU, Y.; NIU, W. A no-reference video quality assessment metric based on ROI. In: LARABI, M.-C.; TRIANTAPHILLIDOU, S. (Ed.). *Image Quality and System Performance XII*. SPIE, 2015. v. 9396, p. 314 – 327. Disponível em: <https://doi.org/10.1117/12.2083892>.

27 Izumi, K.; Kawamura, K.; Yoshino, T.; Naito, S. No reference video quality assessment based on parametric analysis of hevc bitstream. In: *2014 Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*. [S.l.: s.n.], 2014. p. 49–50. ISSN null.

28 Eskicioglu, A. M.; Fisher, P. S. Image quality measures and their performance. *IEEE Transactions on Communications*, v. 43, n. 12, p. 2959–2965, Dec 1995. ISSN 1558-0857.

29 60, I.-T. S. S. P. . C. *Final report from the Video Quality Experts Group on the validation of objective models of video quality assessment, phase II (FR-TV2)*. 2018. Disponível em: <https://www.itu.int/md/T01-SG09-C-0060>.

30 Zhang, W.; Borji, A.; Wang, Z.; Le Callet, P.; Liu, H. The application of visual saliency models in objective image quality assessment: A statistical evaluation. *IEEE Transactions on Neural Networks and Learning Systems*, v. 27, n. 6, p. 1266–1278, June 2016. ISSN 2162-2388.

31 Zhang, W.; Liu, H. Study of saliency in objective video quality assessment. *IEEE Transactions on Image Processing*, v. 26, n. 3, p. 1275–1288, March 2017. ISSN 1941-0042.

32 Feng, X.; Liu, T.; Yang, D.; Wang, Y. Saliency inspired full-reference quality metrics for packet-loss-impaired video. *IEEE Transactions on Broadcasting*, v. 57, n. 1, p. 81–88, March 2011. ISSN 1557-9611.

33 Ćulibrk, D.; Mirković, M.; Zlokolica, V.; Pokrić, M.; Crnojević, V.; Kukolj, D. Salient motion features for video quality assessment. *IEEE Transactions on Image Processing*, v. 20, n. 4, p. 948–958, April 2011. ISSN 1941-0042.

34 CHANDLER, D. M. Seven challenges in image quality assessment: past, present, and future research. *ISRN Signal Processing*, Hindawi Publishing Corporation, v. 2013, 2013.

35 SCHETTINI, R.; CORCHS, S. Underwater image processing: state of the art of restoration and image enhancement methods. *EURASIP Journal on Advances in Signal Processing*, Springer, v. 2010, p. 1–14, 2010.

36 AGAIAN, S. S.; PANETTA, K.; GRIGORYAN, A. M. Transform-based image enhancement algorithms with performance measure. *IEEE Transactions on image processing*, IEEE, v. 10, n. 3, p. 367–382, 2001.

37 ROSENFELD, A. *Digital picture processing*. [S.l.]: Academic press, 1976.

38 UPLAVIKAR, P. M.; WU, Z.; WANG, Z. All-in-one underwater image enhancement using domain-adversarial learning. In: *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*. [S.l.: s.n.], 2019.

39 HAN, F.; YAO, J.; ZHU, H.; WANG, C. Underwater image processing and object detection based on deep cnn method. *Journal of Sensors*, Hindawi, v. 2020, 2020.

40 HOU, W.; GRAY, D. J.; WEIDEMANN, A. D.; FOURNIER, G. R.; FORAND, J. Automated underwater image restoration and retrieval of related optical properties. In: IEEE. *2007 IEEE International Geoscience and Remote Sensing Symposium*. [S.l.], 2007. p. 1889–1892.

41 ABDULLAH-AL-WADUD, M.; KABIR, M. H.; DEWAN, M. A. A.; CHAE, O. A dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, IEEE, v. 53, n. 2, p. 593–600, 2007.

42 HITAM, M. S.; AWALLUDIN, E. A.; YUSSOF, W. N. J. H. W.; BACHOK, Z. Mixture contrast limited adaptive histogram equalization for underwater image enhancement. In: IEEE. *2013 International conference on computer applications technology (ICCAT)*. [S.l.], 2013. p. 1–5.

43 RAHMAN, Z.-u.; JOBSON, D. J.; WOODELL, G. A. Multi-scale retinex for color image enhancement. In: IEEE. *Proceedings of 3rd IEEE International Conference on Image Processing*. [S.l.], 1996. v. 3, p. 1003–1006.

44 WANG, S.; XU, Y.; PANG, Y. A fast underwater optical image segmentation algorithm based on a histogram weighted fuzzy c-means improved by pso. *Journal of Marine Science and Application*, Springer, v. 10, n. 1, p. 70–75, 2011.

45 ANCUTI, C.; ANCUTI, C. O.; HABER, T.; BEKAERT, P. Enhancing underwater images and videos by fusion. In: IEEE. *2012 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.], 2012. p. 81–88.

46 JAFFE, J. S. Computer modeling and the design of optimal underwater imaging systems. *IEEE Journal of Oceanic Engineering*, IEEE, v. 15, n. 2, p. 101–111, 1990.

47 SUBBARAO, M.; LU, M.-C. Image sensing model and computer simulation for ccd camera systems. *Machine Vision and Applications*, Springer, v. 7, n. 4, p. 277–289, 1994.

48 TRUCCO, E.; OLMOS-ANTILLON, A. T. Self-tuning underwater image restoration. *IEEE Journal of Oceanic Engineering*, IEEE, v. 31, n. 2, p. 511–519, 2006.

49 WANG, Y.; ZHANG, J.; CAO, Y.; WANG, Z. A deep cnn method for underwater image enhancement. In: IEEE. *2017 IEEE International Conference on Image Processing (ICIP)*. [S.l.], 2017. p. 1382–1386.

50 LI, C.; GUO, J.; GUO, C. Emerging from water: Underwater image color correction based on weakly supervised color transfer. *IEEE Signal processing letters*, IEEE, v. 25, n. 3, p. 323–327, 2018.

51 PENG, Y.-T.; COSMAN, P. C. Underwater image restoration based on image blurriness and light absorption. *IEEE transactions on image processing*, IEEE, v. 26, n. 4, p. 1579–1594, 2017.

52 ANCUTI, C. O.; ANCUTI, C.; VLEESCHOUWER, C. D.; BEKAERT, P. Color balance and fusion for underwater image enhancement. *IEEE Transactions on image processing*, IEEE, v. 27, n. 1, p. 379–393, 2017.

53 GHANI, A. S. A.; ISA, N. A. M. Underwater image quality enhancement through integrated color model with rayleigh distribution. *Applied soft computing*, Elsevier, v. 27, p. 219–230, 2015.

54 GU, K.; ZHOU, J.; QIAO, J.-F.; ZHAI, G.; LIN, W.; BOVIK, A. C. No-reference quality assessment of screen content pictures. *IEEE Transactions on Image Processing*, IEEE, v. 26, n. 8, p. 4005–4018, 2017.

55 MA, K.; DUANMU, Z.; WU, Q.; WANG, Z.; YONG, H.; LI, H.; ZHANG, L. Waterloo exploration database: New challenges for image quality assessment models. *IEEE Transactions on Image Processing*, IEEE, v. 26, n. 2, p. 1004–1016, 2016.

56 Yang, M.; Sowmya, A. An underwater color image quality evaluation metric. *IEEE Transactions on Image Processing*, v. 24, n. 12, p. 6062–6071, 2015.

57 YANG, M.; SOWMYA, A. New image quality evaluation metric for underwater video. *Signal Processing Letters, IEEE*, v. 21, p. 1215–1219, 10 2014.

58 PANETTA, K.; GAO, C.; AGAIAN, S. Human-visual-system-inspired underwater image quality measures. *IEEE Journal of Oceanic Engineering*, IEEE, v. 41, n. 3, p. 541–551, 2015.

59 CHEN, K.-T.; WU, C.-C.; CHANG, Y.-C.; LEI, C.-L. A crowdsourceable qoe evaluation framework for multimedia content. In: *Proceedings of the 17th ACM international conference on Multimedia*. [S.l.: s.n.], 2009. p. 491–500.

60 IQBAL, K.; SALAM, R. A.; OSMAN, A.; TALIB, A. Z. Underwater image enhancement using an integrated colour model. *IAENG International Journal of computer science*, v. 34, n. 2, 2007.

61 FREITAS, P. G.; ALAMGEER, S.; AKAMINE, W. Y.; FARIAS, M. C. Blind image quality assessment based on multiscale salient local binary patterns. In: *Proceedings of the 9th ACM Multimedia Systems Conference*. [S.l.: s.n.], 2018. p. 52–63.

62 IRSHAD, M.; SANCHEZ-FERREIRA, C.; ALAMGEER, S.; LLANOS, C. H.; FARIAS, M. C. No-reference image quality assessment of underwater images using multi-scale salient local binary patterns. *Electronic Imaging*, Society for Imaging Science and Technology, v. 2021, n. 9, p. 265–1, 2021.

63 ZHANG, L.; ZHANG, L.; MOU, X.; ZHANG, D. Fsim: A feature similarity index for image quality assessment. *IEEE transactions on Image Processing*, IEEE, v. 20, n. 8, p. 2378–2386, 2011.

64 BT, R. I.-R. Methodology for the subjective assessment of the quality of television pictures. *International Telecommunication Union*, 2002.

65 BT, I.-R. R. Subjective assessment methods for image quality in high-definition television. *Question*, v. 211, 1998.

66 RECOMMENDATION, I. P. 910,"subjective video quality assessment methods for multimedia applications,". *International Telecommunication Union, Tech. Rep*, 2008.

67 HOSSFELD, T.; KEIMEL, C.; TIMMERER, C. Crowdsourcing quality-of-experience assessments. *Computer*, IEEE, v. 47, n. 9, p. 98–102, 2014.

68 EILERTSEN, G.; UNGER, J.; MANTIUK, R. K. Evaluation of tone mapping operators for hdr video. In: *High dynamic range video*. [S.l.]: Elsevier, 2016. p. 185–207.

69 AK, A.; GOSWAMI, A.; CALLET, P. L.; DUFAUX, F. A comprehensive analysis of crowdsourcing for subjective evaluation of tone mapping operators. *Electronic Imaging*, Society for Imaging Science and Technology, v. 2021, n. 9, p. 262–1, 2021.

70 SALAS, Ó. F.; ADZIC, V.; SHAH, A.; KALVA, H. Assessing internet video quality using crowdsourcing. In: *Proceedings of the 2nd ACM international workshop on Crowdsourcing for Multimedia*. [S.l.: s.n.], 2013. p. 23–28.

71 GOSWAMI, A.; AK, A.; HAUSER, W.; CALLET, P. L.; DUFAUX, F. Reliability of crowdsourcing for subjective quality evaluation of tone mapping operators. In: *IEEE International Workshop on Multimedia Signal Processing (MMSP'2021)*. [S.l.: s.n.], 2021.

72 BOSC, E.; PEPION, R.; CALLET, P. L.; KOPPEL, M.; NDJIKI-NYA, P.; PRESSIGOUT, M.; MORIN, L. Towards a new quality metric for 3-d synthesized view assessment. *IEEE Journal of Selected Topics in Signal Processing*, IEEE, v. 5, n. 7, p. 1332–1343, 2011.

73 ITU-T. Recommendation P.910: Subjective Video Quality Assessment Methods for Multimedia Applications. *Int. Telecomm. Union, Geneva*, 2008.

74 ITU-R. Recommendation 500-11: Methodology for the Subjective Assessment of the Quality of Television Pictures. *Int. Telecomm. Union, Geneva*, 2002.

75 XU, Q.; HUANG, Q.; YAO, Y. Online crowdsourcing subjective image quality assessment. In: *Proceedings of the 20th ACM international conference on Multimedia*. [S.l.: s.n.], 2012. p. 359–368.

76 GARDLO, B.; EGGER, S.; SEUFERT, M.; SCHATZ, R. Crowdsourcing 2.0: Enhancing execution speed and reliability of web-based qoe testing. In: IEEE. *2014 IEEE International Conference on Communications (ICC)*. [S.l.], 2014. p. 1070–1075.

77 GRZYWALSKI, T.; ŁUCZAK, A.; STASINSKI, R. Internet based subjective assessment of image quality experiment. In: IEEE. *2011 18th International Conference on Systems, Signals and Image Processing*. [S.l.], 2011. p. 1–4.

78 RIBEIRO, F.; FLORENCIO, D.; NASCIMENTO, V. Crowdsourcing subjective image quality evaluation. In: IEEE. *2011 18th IEEE International Conference on Image Processing*. [S.l.], 2011. p. 3097–3100.

79 GHADIYARAM, D.; BOVIK, A. C. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing*, IEEE, v. 25, n. 1, p. 372–387, 2015.

80 MITTAL, A.; MOORTHY, A. K.; BOVIK, A. C.; CHEN, C. W.; CHATZIMISIOS, P.; DAGIUKLAS, T.; ATZORI, L. No-reference approaches to image and video quality assessment. *Multimedia Quality of Experience (QoE): Current Status and Future Requirements*, Wiley Online Library, v. 99, 2015.

81 LAHOULOU, A.; BOURIDANE, A.; VIENNET, E.; HADDADI, M. Full-reference image quality metrics performance evaluation over image quality databases. *Arabian Journal for Science and Engineering*, Springer, v. 38, n. 9, p. 2327–2356, 2013.

82 LIN, W.; KUO, C.-C. J. Perceptual visual quality metrics: A survey. *Journal of visual communication and image representation*, Elsevier, v. 22, n. 4, p. 297–312, 2011.

83 WANG, Z.; BOVIK, A. C.; SHEIKH, H. R.; SIMONCELLI, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, IEEE, v. 13, n. 4, p. 600–612, 2004.

84 CHANDLER, D. M.; HEMAMI, S. S. Vsnr: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE transactions on image processing*, IEEE, v. 16, n. 9, p. 2284–2298, 2007.

85 PONOMARENKO, N.; LUKIN, V.; ZELENSKY, A.; EGIAZARIAN, K.; CARLI, M.; BATTISTI, F. Tid2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of Modern Radioelectronics*, v. 10, n. 4, p. 30–45, 2009.

86 Chandler, D. M.; Hemami, S. S. Vsnr: A wavelet-based visual signal-to-noise ratio for natural images. *IEEE Transactions on Image Processing*, v. 16, n. 9, p. 2284–2298, Sep. 2007. ISSN 1941-0042.

87 Sheikh, H. R.; Bovik, A. C. Image information and visual quality. *IEEE Transactions on Image Processing*, v. 15, n. 2, p. 430–444, Feb 2006. ISSN 1941-0042.

88 SARA, U.; AKTER, M.; UDDIN, M. S. Image quality assessment through fsim, ssim, mse and psnr—a comparative study. *Journal of Computer and Communications*, Scientific Research Publishing, v. 7, n. 3, p. 8–18, 2019.

89 GONZALEZ, R. C. *Digital image processing*. [S.l.]: Pearson education india, 2009.

90 LARSON, E. C.; CHANDLER, D. M. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, International Society for Optics and Photonics, v. 19, n. 1, p. 011006, 2010.

91 Zhang, L.; Lin, W. Introduction to visual attention. In: ____. *Selective Visual Attention: Computational Models and Applications*. [S.l.]: IEEE, 2013. p. 1–24. ISBN 9781118060056978047082813 7.

92 Itti, L.; Koch, C.; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, v. 20, n. 11, p. 1254–1259, Nov 1998. ISSN 1939-3539.

93 HAREL, J.; KOCH, C.; PERONA, P. Graph-based visual saliency. In: *Proceedings of the 19th International Conference on Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2006. (NIPS'06), p. 545–552. Disponível em: <http://dl.acm.org/citation.cfm?id=2976456.2976525>.

94 Sadaka, N. G.; Karam, L. J.; Ferzli, R.; Abousleman, G. P. A no-reference perceptual image sharpness metric based on saliency-weighted foveal pooling. In: *2008 15th IEEE International Conference on Image Processing*. [S.l.: s.n.], 2008. p. 369–372.

95 Rao, D. V.; Sudhakar, N.; Babu, I. R.; Reddy, L. P. Image quality assessment complemented with visual regions of interest. In: *2007 International Conference on Computing: Theory and Applications (ICCTA'07)*. [S.l.: s.n.], 2007. p. 681–687.

96 ZHANG, L.; SHEN, Y.; LI, H. Vsi: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, IEEE, v. 23, n. 10, p. 4270–4281, 2014.

97 FARIAS, M. C.; AKAMINE, W. Y. On performance of image quality metrics enhanced with visual attention computational models. *Electronics letters*, IET, v. 48, n. 11, p. 631–633, 2012.

98 ENGELKE, U.; KAPRYKOWSKY, H.; ZEPERNICK, H.-J.; NDJIKI-NYA, P. Visual attention in quality assessment. *IEEE Signal Processing Magazine*, IEEE, v. 28, n. 6, p. 50–59, 2011.

99 GU, K.; WANG, S.; YANG, H.; LIN, W.; ZHAI, G.; YANG, X.; ZHANG, W. Saliency-guided quality assessment of screen content images. *IEEE Transactions on Multimedia*, IEEE, v. 18, n. 6, p. 1098–1110, 2016.

100 Zhang, L.; Lin, W. Application of attention models in image processing. In: ____. *Selective Visual Attention: Computational Models and Applications*. [S.l.]: IEEE, 2013. p. 271–303. ISBN 9781118060056978047082813 7.

101 LI, J.; ZHOU, Y. Visual saliency based blind image quality assessment via convolutional neural network. Springer International Publishing, p. 550–557, 2017.

102 ALGINA, H. J. K. J. Comparing squared multiple correlation coefficients: Examination of a confidence interval and a test significance. *Psychological Methods*, n. 1939-1463, p. 76–83, 1999.

103 BOBKO, P. *Correlation and regression: Applications for industrial organizational psychology and management (2nd ed.).* [S.l.]: CA: Sage Publications, 2001.

104 YANG, Z.; DAI, Z.; YANG, Y.; CARBONELL, J.; SALAKHUTDINOV, R. R.; LE, Q. V. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, v. 32, 2019.

105 LU, L.; GUO, X.; ZHAO, J. A unified nonlocal strain gradient model for nanobeams and the importance of higher order terms. *International Journal of Engineering Science*, Elsevier, v. 119, p. 265–277, 2017.

106 ZHANG, S.; YAO, L.; SUN, A.; TAY, Y. Deep learning based recommender system: A survey and new perspectives. *ACM Computing Surveys (CSUR)*, ACM New York, NY, USA, v. 52, n. 1, p. 1–38, 2019.

107 LI, C.; ANWAR, S.; PORIKLI, F. Underwater scene prior inspired deep underwater image and video enhancement. *Pattern Recognition*, Elsevier, v. 98, p. 107038, 2020.

108 KUMAR, S.; STECHER, G.; LI, M.; KNYAZ, C.; TAMURA, K. Mega x: molecular evolutionary genetics analysis across computing platforms. *Molecular biology and evolution*, Oxford University Press, v. 35, n. 6, p. 1547, 2018.

109 MAINI, R.; AGGARWAL, H. A comprehensive review of image enhancement techniques. *arXiv preprint arXiv:1003.4053*, 2010.

110 HE, J.; OSSENBRUGGEN, J. V.; VRIES, A. de. Fish4label: accomplishing an expert task without expert knowledge. In: . [S.l.: s.n.], 2013. p. 211–212.

111 XIAO, J.; EHINGER, K. A.; HAYS, J.; TORRALBA, A.; OLIVA, A. Sun database: Exploring a large collection of scene categories. *Int. J. Comput. Vision*, Kluwer Academic Publishers, USA, v. 119, n. 1, p. 3–22, ago. 2016. ISSN 0920-5691. Disponível em: <https://doi.org/10.1007/s11263-014-0748-y>.

112 (MARIS), T. I. M. I. S. *The Marine Information System*. 2019. Disponível em: <https://maris.iaea.org/home>.

113 AKKAYNAK, D.; TREIBITZ, T. Sea-thru: A method for removing water from underwater images. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2019. p. 1682–1691.

114 ARAVIND, P. *Aravind Pai' answer to Ẅhy is deep learning called as such?¨*. 2018. Disponível em: <https://www.quora.com/Why-is-deep-learning-called-as-such>.

115 ALPAYDIN, E. *Introduction to Machine Learning*. [S.l.]: The MIT Press, 2014. ISBN 0262028182, 9780262028189.

116 YE J. KUMAR, L. K. P.; DOERMANN, D. Unsupervised feature learning framework for no-reference image quality assessment. *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012.

117 LIU, L.; DONG, H.; HUANG, H.; BOVIK, A. C. No-reference image quality assessment in curvelet domain. *Signal Processing: Image Communication*, Elsevier, v. 29, n. 4, p. 494–505, 2014.

118 LIU B. LIU, H. H. L.; BOVIK, A. No-reference image quality assessment based on spatial and spectral entropies. *Signal Processing: Image Communication*, June 2014.

119 MITTAL, A. K. M. A.; BOVIK, A. C. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 2012.

120 FREITAS, P. G.; AKAMINE, W. Y.; FARIAS, M. C. No-reference image quality assessment based on statistics of local ternary pattern. In: IEEE. *Quality of Multimedia Experience (QoMEX), 2016 Eighth International Conference on*. [S.l.], 2016. p. 1–6.

121 MOORTHY, A. K.; BOVIK, A. C. Blind image quality assessment: From scene statistics to perceptual quality. *EEE Transactions Image Processing*, p. 3350–3364, December 2011.

122 Li, Q.; Lin, W.; Fang, Y. No-reference quality assessment for multiply-distorted images in gradient domain. *IEEE Signal Processing Letters*, v. 23, n. 4, p. 541–545, April 2016. ISSN 1070-9908.

123 Saad, M. A.; Bovik, A. C.; Charrier, C. Blind prediction of natural video quality. *IEEE Transactions on Image Processing*, v. 23, n. 3, p. 1352–1365, March 2014. ISSN 1057-7149.

124 Li, X.; Guo, Q.; Lu, X. Spatiotemporal statistics for video quality assessment. *IEEE Transactions on Image Processing*, v. 25, n. 7, p. 3329–3342, July 2016. ISSN 1057-7149.

125 FREITAS, P. G.; AKAMINE, W. Y.; FARIAS, M. C. Using multiple spatio-temporal features to estimate video quality. *Signal Processing: Image Communication*, v. 64, p. 1 – 10, 2018. ISSN 0923-5965.

126 MEN, H.; LIN, H.; SAUPE, D. Spatiotemporal feature combination model for no-reference video quality assessment. *IEEE*, 04 2018.

127 SINGH, R.; AGGARWAL, N. A distortion-agnostic video quality metric based on multi-scale spatio-temporal structural information. *Signal Processing: Image Communication*, v. 74, p. 299 – 308, 2019. ISSN 0923-5965.

128 LECUN, Y.; BOTTOU, L.; BENGIO, Y.; HAFFNER, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, Ieee, v. 86, n. 11, p. 2278–2324, 1998.

129 NIU, X.-X.; SUEN, C. Y. A novel hybrid cnn–svm classifier for recognizing handwritten digits. *Pattern Recognition*, Elsevier, v. 45, n. 4, p. 1318–1325, 2012.

130 RUSSAKOVSKY, O.; DENG, J.; SU, H.; KRAUSE, J.; SATHEESH, S.; MA, S.; HUANG, Z.; KARPATHY, A.; KHOSLA, A.; BERNSTEIN, M. et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, Springer, v. 115, n. 3, p. 211–252, 2015.

131 ZEILER, M. D.; FERGUS, R. Visualizing and understanding convolutional networks. In: SPRINGER. *European conference on computer vision*. [S.l.], 2014. p. 818–833.

132 SIMONYAN, K.; ZISSERMAN, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

133 SZEGEDY, C.; LIU, W.; JIA, Y.; SERMANET, P.; REED, S.; ANGUELOV, D.; ERHAN, D.; VANHOUCKE, V.; RABINOVICH, A. Going deeper with convolutions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2015. p. 1–9.

134 HE, K.; ZHANG, X.; REN, S.; SUN, J. Deep residual learning for image recognition. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. [S.l.: s.n.], 2016. p. 770–778.

135 Kang, L.; Ye, P.; Li, Y.; Doermann, D. Convolutional neural networks for no-reference image quality assessment. In: *2014 IEEE Conference on Computer Vision and Pattern Recognition*. [S.l.: s.n.], 2014. p. 1733–1740. ISSN 1063-6919.

136 DOMONKOS, V. No-reference video quality assessment based on the temporal pooling of deep features. In: . [S.l.: s.n.], 2019.

137 Ahn, S.; Lee, S. Deep blind video quality assessment based on temporal human perception. In: *2018 25th IEEE International Conference on Image Processing (ICIP)*. [S.l.: s.n.], 2018. p. 619–623.

138 Kim, J.; Lee, S. Fully deep blind image quality predictor. *IEEE Journal of Selected Topics in Signal Processing*, v. 11, n. 1, p. 206–220, Feb 2017.

139 DOMONKOS, V.; SZIRÁNYI, T. No-reference video quality assessment via pretrained cnn and lstm networks. *Signal, Image and Video Processing*, Jun 2019.

140 KIM, D.; KIM, J.; KIM, J. Elastic exponential linear units for convolutional neural networks. *Neurocomputing*, v. 406, p. 253–266, 2020. ISSN 0925-2312. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0925231220304240>.

141 SUTSKEVER, I.; MARTENS, J.; DAHL, G.; HINTON, G. On the importance of initialization and momentum in deep learning. In: *Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28*. [S.l.]: JMLR.org, 2013. (ICML'13), p. III–1139–III–1147.

142 Swasono, D. I.; Tjandrasa, H.; Fathicah, C. Classification of tobacco leaf pests using vgg16 transfer learning. In: *2019 12th International Conference on Information Communication Technology and System (ICTS)*. [S.l.: s.n.], 2019. p. 176–181. ISSN null.

143 VASUDEV, R. *Understanding and Calculating the number of Parameters in Convolution Neural Networks (CNNs)*. 2011. Disponível em: <https://towardsdatascience.com/understanding-and-calculating-the-number-of-parameters-in-convolution-neural-networks-cnns-fc88790d530d>.

144 SAHA, S. *A Comprehensive Guide to Convolutional Neural Networks — the ELI5 way*. 2018. Disponível em: <https://towardsdatascience.com/a-comprehensive-guide-to-convolutional-neural-networks-the-eli5-way-3bd2b1164a53>.

145 VU, C.; PHAN, T.; BANGA, P.; CHANDLER, D. On the quality assessment of enhanced images: A database, analysis, and strategies for augmenting existing methods. In: . [S.l.: s.n.], 2012. p. 181–184. ISBN 978-1-4673-1831-0.

146 RAMPONI, G. A cubic unsharp masking technique for contrast enhancement. *Signal Processing*, Elsevier, v. 67, n. 2, p. 211–222, 1998.

147 YOUNG, I. T.; VLIET, L. J. V. Recursive implementation of the gaussian filter. *Signal processing*, Elsevier, v. 44, n. 2, p. 139–151, 1995.

148 Abdullah-Al-Wadud, M.; Kabir, M. H.; Akber Dewan, M. A.; Chae, O. A dynamic histogram equalization for image contrast enhancement. *IEEE Transactions on Consumer Electronics*, v. 53, n. 2, p. 593–600, May 2007. ISSN 1558-4127.

149 SONG, K. S.; KANG, H.; KANG, M. G. Hue-preserving and saturation-improved color histogram equalization algorithm. *JOSA A*, Optical Society of America, v. 33, n. 6, p. 1076–1088, 2016.

150 REN, X.; LI, M.; CHENG, W.-H.; LIU, J. Joint enhancement and denoising method via sequential decomposition. 04 2018.

151 KEELAN, B. *Handbook of image quality: characterization and prediction*. [S.l.]: CRC Press, 2002.

152 PINSON, M. H.; JANOWSKI, L.; PAPIR, Z. Video quality assessment: subjective testing of entertainment scenes. *IEEE Signal Processing Magazine*, IEEE, v. 32, n. 1, p. 101–114, 2014.

153 ITU-T. Recommendation P.913: Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment. *Int. Telecomm. Union, Geneva*, 2016.

154 YING, Z.; LI, G.; REN, Y.; WANG, R.; WANG, W. A new image contrast enhancement algorithm using exposure fusion framework. In: . [S.l.: s.n.], 2017. p. 36–46. ISBN 978-3-319-64697-8.

155 LUNDH, F. *Pillow: The friendly PIL fork*. 2019. Disponível em: <https://python-pillow.org/>.

156 MIKE. *Enhancing Photos with Python*. 2017. Disponível em: <http://www.blog.pythonlibrary.org/2017/10/24/enhancing-photos-with-python/>.

157 MATHWORKS. *Image Filtering and Enhancement*. 2019. Disponível em: <https://www.mathworks.com/help/images/image-enhancement-and-restoration.html?s_tid=CRUX_lftnav>.

158 Hoßfeld, T.; Keimel, C.; Hirth, M.; Gardlo, B.; Habigt, J.; Diepold, K.; Tran-Gia, P. Best practices for qoe crowdtesting: Qoe assessment with crowdsourcing. *IEEE Transactions on Multimedia*, v. 16, n. 2, p. 541–558, Feb 2014. ISSN 1941-0077.

159 FREITAS, P. G.; ALAMGEER, S.; AKAMINE, W. Y. L.; FARIAS, M. C. Q. Blind image quality assessment based on multiscale salient local binary patterns. In: . New York, NY, USA: Association for Computing Machinery, 2018. (MMSys '18), p. 52–63. ISBN 9781450351928. Disponível em: <https://doi.org/10.1145/3204949.3204960>.

160 OJALA, T.; PIETIKÄINEN, M.; MÄENPÄÄ, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, v. 24, p. 971–987, 2002.

161 ZHANG, J.; SCLAROFF, S. Exploiting surroundedness for saliency detection: a boolean map approach. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 38, n. 5, p. 889–902, 2016.

162 FERNÁNDEZ-DELGADO, M.; CERNADAS, E.; BARRO, S.; AMORIM, D. Do we need hundreds of classifiers to solve real world classification problems. *Journal of Machine Learning Research*, v. 15, n. 1, p. 3133–3181, 2014.

163 Mittal, A.; Soundararajan, R.; Bovik, A. C. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, v. 20, n. 3, p. 209–212, 2013.

164 ZHOU, Y.; LI, L.; WANG, S.; JINJIAN, W.; FANG, Y.; GAO, X. No-reference quality assessment for view synthesis using dog-based edge statistics and texture naturalness. *IEEE Transactions on Image Processing*, PP, p. 1–1, 04 2019.

165 BALBOA, R. M.; GRZYWACZ, N. M. Occlusions and their relationship with the distribution of contrasts in natural images. *Vision Research*, v. 40, n. 19, p. 2661 – 2669, 2000. ISSN 0042-6989. Disponível em: <http://www.sciencedirect.com/science/article/pii/S0042698900000997>.

166 ZHANG, L.; GU, Z.; LI, H. Sdsp: A novel saliency detection method by combining simple priors. *2013 IEEE International Conference on Image Processing*, p. 171–175, 2013.

167 HARRIS, C.; STEPHENS, M. A combined corner and edge detector. In: *Alvey Vision Conference*. [S.l.: s.n.], 1988.

168 SHRIVAKSHAN, G.; CHANDRASEKAR, C. A comparison of various edge detection techniques used in image processing. *International Journal of Computer Science Issues*, v. 9, p. 269–276, 09 2012.

169 ALAMGEER, S.; IRSHAD, M.; FARIAS, M. C. Cnn-based no-reference video quality assessment method using a spatiotemporal saliency patch selection procedure. *Journal of Electronic Imaging*, SPIE, v. 30, n. 6, p. 063001, 2021.

170 JAFFE, J. S. Underwater optical imaging: the past, the present, and the prospects. *IEEE Journal of Oceanic Engineering*, IEEE, v. 40, n. 3, p. 683–700, 2014.

171 LI, C.; GUO, C.; REN, W.; CONG, R.; HOU, J.; KWONG, S.; TAO, D. An underwater image enhancement benchmark dataset and beyond. *IEEE Transactions on Image Processing*, v. 29, p. 4376–4389, 2020.

172 KADYROVA, A.; PEDERSEN, M.; AHMAD, B.; MANDAL, D. J.; NGUYEN, M.; ZIMMERMANN, P. H. Image enhancement dataset for evaluation of image quality metrics. *Electronic Imaging*, Society for Imaging Science and Technology, v. 34, p. 1–6, 2022.

173 FU, X.; FAN, Z.; LING, M.; HUANG, Y.; DING, X. Two-step approach for single underwater image enhancement. In: IEEE. *2017 international symposium on intelligent signal processing and communication systems (ISPACS)*. [S.l.], 2017. p. 789–794.

174 FU, X.; ZHUANG, P.; HUANG, Y.; LIAO, Y.; ZHANG, X.-P.; DING, X. A retinex-based enhancing approach for single underwater image. In: IEEE. *2014 IEEE international conference on image processing (ICIP)*. [S.l.], 2014. p. 4572–4576.

175 DREWS, P. L.; NASCIMENTO, E. R.; BOTELHO, S. S.; CAMPOS, M. F. M. Underwater depth estimation and image restoration based on single images. *IEEE computer graphics and applications*, IEEE, v. 36, n. 2, p. 24–35, 2016.

176 LI, C.; GUO, J.; GUO, C.; CONG, R.; GONG, J. A hybrid method for underwater image correction. *Pattern Recognition Letters*, Elsevier, v. 94, p. 62–67, 2017.

177 PENG, Y.-T.; CAO, K.; COSMAN, P. C. Generalization of the dark channel prior for single image restoration. *IEEE Transactions on Image Processing*, IEEE, v. 27, n. 6, p. 2856–2868, 2018.

178 GALDRAN, A.; PARDO, D.; PICÓN, A.; ALVAREZ-GILA, A. Automatic red-channel underwater image restoration. *Journal of Visual Communication and Image Representation*, Elsevier, v. 26, p. 132–145, 2015.

179 LI, C.-Y.; GUO, J.-C.; CONG, R.-M.; PANG, Y.-W.; WANG, B. Underwater image enhancement by dehazing with minimum information loss and histogram distribution prior. *IEEE Transactions on Image Processing*, IEEE, v. 25, n. 12, p. 5664–5677, 2016.

180 HE, K.; SUN, J.; TANG, X. Single image haze removal using dark channel prior. *IEEE transactions on pattern analysis and machine intelligence*, IEEE, v. 33, n. 12, p. 2341–2353, 2010.

181 REN, W.; LIU, S.; ZHANG, H.; PAN, J.; CAO, X.; YANG, M.-H. Single image dehazing via multi-scale convolutional neural networks. In: SPRINGER. *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*. [S.l.], 2016. p. 154–169.

182 WANG, Z.; BOVIK, A. C.; SHEIKH, H. R.; SIMONCELLI, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, IEEE, v. 13, n. 4, p. 600–612, 2004.

183 KINGMA D., B. J. A. A method for stochastic optimization. *arXiv preprint*, 2014.

184 ZHANG, W.; LIU, H. Study of saliency in objective video quality assessment. *IEEE Transactions on Image Processing*, IEEE, v. 26, n. 3, p. 1275–1288, 2017.

185 ZHANG, W.; BORJI, A.; WANG, Z.; CALLET, P. L.; LIU, H. The application of visual saliency models in objective image quality assessment: A statistical evaluation. *IEEE transactions on neural networks and learning systems*, IEEE, v. 27, n. 6, p. 1266–1278, 2015.

186 ALAMGEER, S.; FARIAS, M. A two-stream cnn based visual quality assessment method for light field images. *Multimedia Tools and Applications*, jul 2022.

187 ALAMGEER, S.; FARIAS, M. No-reference light field image quality assessment method based on a long-short term memory neural network. *IEEE INTERNATIONAL CONFERENCE ON MULTIMEDIA AND EXPO (ICME), Taipei, Taiwan*, 2022.

188 ALAMGEER, S.; FARIAS, M. Deep learning-based light field image quality assessment using frequency domain inputs. *The 14th International Conference on Quality of Multimedia Experience (QoMEX), Lippstadt, Germany*, 2022.

Apendix A

PAPERS RESULTING FROM THIS THESIS

Journal Paper

1. Sana Alamgeer, Muhammad Irshad, Mylène C. Q. Farias. CNN-based no-reference video quality assessment method using a spatiotemporal saliency patch selection procedure. https://DOI: 10.1117/1.JEI.30.6.063001

Conference Papers

1. Muhammad Irshad, Alessandro R Silva, Sana Alamgeer, and Mylène C. Q. Farias. Perceptual Quality Assessment of Enhanced Images Using a Crowd-Sourcing Framework. https://doi.org/10.2352/ISSN.2470-1173.2020.9.IQSP-066

2. Muhammad Irshad, Camilo Sanchez Ferreira, Sana Alamgeer, Carlos H. Llanos, Mylène C.Q. Farias. No-reference Image Quality Assessment of Underwater Images Using Multi-Scale Salient Local Binary Patterns. https://doi.org/10.2352/ISSN.2470-1173.2021.9.IQSP-265

Papers Under Review

1. Muhammad Irshad, Camilo Sanchez-Ferreira, Sana Alamgeer, Carlos H. Llano, Mylène C.Q. Farias. No-Reference Underwater Image Quality Assessment based on Convolutional Neural Network.

First Page of Published Papers

https://doi.org/10.2352/ISSN.2470-1173.2020.9.IQSP-066

Perceptual Quality Assessment of Enhanced Images Using a Crowd-Sourcing Framework

***Muhammad Irshad*[1], *Alessandro R. Silva*[2], *Sana Alamgeer*[1], *Mylène C.Q. Farias*[1];**
[1] ***Department of Electrical Engineering,*** [2] ***Department of Computer Science,***
University of Brasilia, Brazil.

Abstract

In this work, we present a psychophysical study, in which, we analyzed the perceptual quality of images enhanced with several types of enhancement algorithms, including color, sharpness, histogram, and contrast enhancements. To estimate and compare the qualities of enhanced images, we performed a psychophysical experiment with 35 source images, obtained from publicly available databases. More specifically, we used images from the Challenge Database, the CSIQ database, and the TID2013 database. To generate the test sequences, we used 12 different image enhancement algorithms, generating a dataset with a total of 455 images. We used a Double Stimulus Continuous Quality Scale (DSCQS) experimental methodology, with a between-subjects approach where each subject scored a subset of the total database to avoid fatigue. Given the high number of test images, we designed a crowd-sourcing interface to perform an online psychophysical experiment. This type of interface has the advantage of making it possible to collect data from many participants. We also performed an experiment in a controlled laboratory environment and compared its results with the crowd-sourcing results. Since there are very few quality enhancement databases available in the literature, this works represents a contribution to the area of image quality.

Keywords: Enhancement; Perceptual Quality Assessment; Crowd-Sourcing Framework, Subjective Quality Assessment.

Introduction

Image enhancement is frequently used to improve or restore the visual quality of images and videos. Currently, there are several image enhancement algorithms, but there is not yet a performance metric that is able to estimate the performance of these methods. Since the final consumers of the resulting enhanced visual content are human viewers, the performance of these algorithms should be measured by estimating the visual quality of the enhanced images, taking into consideration the human visual system [20].

Image quality can be estimated using subjective (psychophysical experiments) and objective (quality metrics) methods [9, 21]. Subjective methods are simply psychophysical experiments where participants rate one or more aspects of a set of processed images. Most often, these experiments are performed in a controlled environment (e.g. a laboratory), following standard recommendations for the environment conditions and experimental methodologies [6]. It worth pointing out that although data (subjective scores) collected in psychophysical experiments are considered as ground-truth, these experiments are time-consuming and expensive. Objective quality methods, on the other hand, are algorithms (implemented in hardware or software) that automatically estimate the quality of an image [14, 10]. These methods are designed and tested using subjective scores as ground-truth.

The area of image and video quality has achieved great progress in the last decades [2]. But, although the performance accuracy of quality metrics has improved, there are still many challenges in this area. Among them is the design of objective quality metrics for enhanced contents. Since most of the quality metrics have been designed to capture visual distortions, they are not able to quantify the changes in quality introduced by enhancement algorithms. Therefore, currently, there is a need for quality metrics that can automatically estimate the quality of enhanced images and videos. It is worth pointing out that developing quality metrics for enhanced images is a challenge due to the lack of quality databases containing enhanced images and their respective (ground-truth) subjective quality scores.

In this paper, our goal is to introduce a quality database for enhanced images. Up to our knowledge, currently, there is only one image enhancement quality database that can be used for research in image quality [19]. However, this database contains images of low resolution that were processed manually, using a professional graphics editing software (Adobe Photoshop) to produce the best possible enhanced images. In our database, we used images of a higher resolution, which are enhanced with twelve different image enhancement algorithms. Our goal was to produce a set of images that were like consumer applications contents. Also, we performed a crowd-sourcing subjective experiment to obtain quality scores for all database images. With this experiment, we were able to obtain a large and diverse pool of participants.

Database Content Generation

Figure 1 shows a block diagram of the strategy used to generate the database. Our first step was to choose 35 original (source - SRC) images. These images were taken from three image quality databases, to allow for future comparisons of enhanced and degraded images. More specifically, we took 5 SRC images from the CSIQ database [18], 5 original images from the TID2013 database [16], and 25 original images from the ChallengeDB database [3]. Table 6 (in the Appendix) shows a list of the SRC images, along with their names in the corresponding databases. These chosen source contents are diverse, in terms of spatial activity, semantic content, and color distribution. In Figure 2, the first row (SRCs) shows examples of SRC images taken from the (a-b) TID2013, (c-d) CSIQ, and (e-f) ChallengeDB databases.

Our next step consists of choosing the enhancement algo-

https://doi.org/10.2352/ISSN.2470-1173.2021.9.IQSP-265

No-reference Image Quality Assessment of Underwater Images Using Multi-Scale Salient Local Binary Patterns

Muhammad Irshad[1], Camilo Sanchez-Ferreira[2], Sana Alamgeer[1], Carlos H. Llanos[3], and Mylène C.Q. Farias[1]

[1]Department of Electrical Engineering, University of Brasília, Brazil.
[2]Department of Physics, University of Cauca, Colombia.
[3]Department of Mechanical Engineering, University of Brasília, Brazil.

Abstract

Images acquired in underwater scenarios may contain severe distortions due to light absorption and scattering, color distortion, poor visibility, and contrast reduction. Because of these degradations, researchers have proposed several algorithms to restore or enhance underwater images. One way to assess these algorithms' performance is to measure the quality of the restored/enhanced underwater images. Unfortunately, since reference (pristine) images are often not available, designing no-reference (blind) image quality metrics for this type of scenario is still a challenge. In fact, although the area of image quality has evolved a lot in the last decades, estimating the quality of enhanced and restored images is still an open problem. In this work, we present a no-reference image quality evaluation metric for enhanced underwater images (NR-UWIQA) that uses an adapted version of the multi-scale salient local binary pattern operator to extract image features and a machine learning approach to predict quality. The proposed metric was tested on the UID-LEIA database and presented good accuracy performance when compared to other state-of-the-art methods. In summary, the proposed NR-UWQIA method can be used to evaluate the results of restoration techniques quickly and efficiently, opening a new perspective in the area of underwater image restoration and quality assessment.

Keywords: Underwater image enhancement; Image quality assessment; Quality metrics, full-reference, no-reference; Underwater image formation model; Saliency; Multiscale Salient Local Binary Patterns;

Introduction

Underwater images are often characterized by a poor visibility since the light travelling in the water medium is attenuated and, consequently, the captured scenes may be poorly contrasted and hazy. More specifically, light attenuation is produced by absorption and scattering processes. Absorption removes the light energy while scattering changes the direction of the light. Therefore, underwater images may have different types of degradations, including limited-range visibility, non-uniform lightening, low contrast, blurring, diminished color, bright artifacts, and noise. In other words, the visual aspect of underwater images may vary a lot depending on the water medium's characteristics, including the types of particles present in the water and the water depth [26]. Figure 1 shows examples of images captured underwater in three different scenarios: shallow water, deep water, and muddy water. Notice that, generally, degradations of images captured underwater are stronger than degradations of images captured over-the-air [39]. Often the quality of underwater images is not adequate for the to be used by image and computer vision algorithms, requiring the use of restoration or enhacement algorithms [39].

Given the importance of the overall quality of underwater-captured images for ocean engineering and scientific research, there are in the literature several methods for restoring or enhancing the quality of underwater images [16, 19]. Therefore, the use of underwater images in computer vision and image processing applications often depends on the success restoration and enhancement algorithms [35, 3, 15]. To determine the performance of these algorithms, we must estimate the quality of the restored/enhanced images as perceived by human viewers. Unfortunately, most methods used to estimate the performance of these algorithms do not consider human perception or image quality. One of the reasons is that subjective quality experiments, which are considered as the ground truth in image quality research, are costly and time-consuming [21]. Moreover, these methods are unfeasible for real-time applications and system integration. One viable option to estimate the quality of restored or enhanced underwater images and, therefore, the restoration algorithm's performance is to use objective image quality assessment (IQA) methods.

IQA methods are algorithms capable of automatically estimate the quality of an image. These methods can be divided into three classes: (a) full-reference (FR) IQA methods, where a reference image is needed to estimate the quality; (2) reduced reference (RR) IQA methods, where partial information about the reference image is available; or (3) no-reference (NR) IQA methods, which blindly estimate quality without having access to the reference or pristine image. For underwater images scenario, where a reference image is not available, we must use NR-IQA methods to estimate the perceptual quality of restored and degraded images. IQA methods can be used to evaluate the restoration process's success and determine if the images are adequate for the target underwater engineering and monitoring applications. So far, a few researchers have proposed IQA methods specifically for underwater images. For example, Sanchez *et al.* [37] have proposed a restoration algorithm for underwater that uses an NR-IQA method as a performance metric for the optimization algorithm.

Although in the last decades a lot of progress has been made in the area of image quality assessment, designing metrics to estimate the quality of enhanced and restored images remains a challenge [8]. As mentioned earlier, the final quality of underwater images depend on the marine habitats where the images are captured, which often introduce specific chroma, saturation, and con-

CNN-based no-reference video quality assessment method using a spatiotemporal saliency patch selection procedure

Sana Alamgeer,* **Muhammad Irshad, and Mylène C. Q. Farias**
University of Brasília, Department of Electrical Engineering, Brasília, Brazil

Abstract. We propose a yet lightweight no-reference (NR) video quality assessment (VQA) method, which uses a convolution neural network (CNN) architecture. The proposed method implements a spatiotemporal saliency patch selection procedure that crops the frame into small nonoverlapping blocks of images (patches) and selects the most perceptually relevant ones. The selected patches are then forwarded to the CNN. To determine which patches are the most relevant, spatial and temporal saliency features are computed for each frame. The proposed method does not require subjective scores to train the CNN. It uses objective quality scores as target quality scores for each video frame, which are computed using an NR image quality assessment method. Given the lack of large annotated video quality databases, this is an advantage of the proposed method. Finally, although it has much smaller cost of data-processing, compared with other state-of-the-art methods, the proposed NR-VQA obtains robust and competitive results. © *2021 SPIE and IS&T* [DOI: 10.1117/1.JEI.30.6.063001]

Keywords: video quality assessment; saliency; convolution neural network; objective quality scores.

Paper 210107 received Feb. 28, 2021; accepted for publication Oct. 13, 2021; published online Nov. 1, 2021.

1 Introduction

In the last decades, there has been a tremendous increase in the popularity of video applications, with 82% of the internet traffic being currently video data.[1] Since the success or popularity of a video service is correlated to the quality of experience of the end user,[2] it is often important to assess the quality of the video signal at the client side, and the quality assessment is performed using quality assessment methods.

Quality assessment methods are algorithms that estimate the quality of videos (VQA) or images (IQA) either objectively or subjectively. Subjective quality assessment methods estimate the quality of images/videos by performing psychophysical experiments, where participants assign a score to each image/video. An estimate of the quality is given by the mean observer score (MOS), which is computed by averaging the scores given to a test image/video by all participants. Although subjective image/video quality assessment (VQA) methods are considered as ground-truth in image/video quality, these methods are expensive and time consuming. Objective image/VQA methods, on the other hand, estimate the quality of image/video using computational algorithms (quality metrics), which are faster, cheaper, and can be more easily incorporated in a multimedia application. In this work, henceforth, we use acronyms IQA and VQA to refer to objective image quality assessment (IQA) methods and video quality assessment methods, respectively.

VQA methods can be classified as full-reference (FR), reduced-reference (RR), and no-reference (NR) methods. FR VQA methods, which require the reference (pristine) content, are frequently the best performing metrics.[3] RR VQA methods require sending (or embedding) features from the original content to the receiver/user,[4] whereas NR VQA methods estimate quality blindly without having access to the original.[5] Unfortunately, both FR and RR metrics cannot be used in real-time applications where the reference or even a small amount of the reference

*Address all correspondence to Sana Alamgeer, sanaalamgeer@gmail.com

1017-9909/2021/$28.00© 2021 SPIE and IS&T