



Progress in Blind Image Quality Assessment: A Brief Review

Pei Yang ¹ , Jordan Sturtz ² and Letu Qingge ^{2,*} 

¹ Department of Computer Technology and Application, Qinghai University, Xining 810016, China; yangpeinmgdx@sina.com

² Department of Computer Science, North Carolina A&T State University, Greensboro, NC 27411, USA; jasturtz@aggies.ncat.edu

* Correspondence: lqingge@ncat.edu

Abstract: As a fundamental research problem, blind image quality assessment (BIQA) has attracted increasing interest in recent years. Although great progress has been made, BIQA still remains a challenge. To better understand the research progress and challenges in this field, we review BIQA methods in this paper. First, we introduce the BIQA problem definition and related methods. Second, we provide a detailed review of the existing BIQA methods in terms of representative hand-crafted features, learning-based features and quality regressors for two-stage methods, as well as one-stage DNN models with various architectures. Moreover, we also present and analyze the performance of competing BIQA methods on six public IQA datasets. Finally, we conclude our paper with possible future research directions based on a performance analysis of the BIQA methods. This review will provide valuable references for researchers interested in the BIQA problem.

Keywords: blind image quality assessment; no-reference image quality assessment; natural scene statistics; mean opinion score; one-stage BIQA

MSC: 68T45



Citation: Yang, P.; Sturtz, J.; Qingge, L. Progress in Blind Image Quality Assessment: A Brief Review. *Mathematics* **2023**, *11*, 2766. <https://doi.org/10.3390/math11122766>

Academic Editor: Samaneh Mazaheri

Received: 26 March 2023

Revised: 3 May 2023

Accepted: 16 June 2023

Published: 19 June 2023



Copyright: © 2023 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

With the ubiquity of image-capture devices and the growth of the internet, digital images are playing an increasingly important role in every aspect of daily life. During the many stages of image handling, including acquisition, transmission, processing, storage, etc., various distortions in image quality may occur. Image quality assessment (IQA) has become an important need in a wide range of computer-vision-related applications, such as image retrieval [1,2] and visual recognition [3,4]. Humans have the ability to subjectively evaluate the quality of digital images. In fact, this subjective assessment is one of the most reliable methods for IQA [5]. However, subjective quality assessment is laborious and time-consuming. In addition, it is difficult to implement it in real-world application scenarios. Thus, considerable effort has been made to develop objective IQA algorithms in the past decades [6–9].

Objective IQA algorithms aim to mimic the capacity of the human vision system (HVS) to assess the quality of images. According to the availability of reference images (i.e., pristine images), we can roughly divide the existing objective IQA methods into full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA), and no-reference or blind IQA (NR-IQA/BIQA) [10]. For the FR-IQA and RR-IQA methods, a reference image is completely or partially used to assess the quality of a distorted image [6,11]. The FR-IQA and RR-IQA methods usually have remarkable performance due to the usage of reference images. However, they are very limited in practical applications as the reference image is often expensive to obtain or completely unavailable. In contrast, the NR-IQA or BIQA methods are more attractive and applicable to various applications since they do not require reference images [12].

In the past decades, many BIQA algorithms have been proposed including distortion-specific and non-distortion-specific methods [13]. Distortion-specific methods predict the quality of images with certain types of distortion (e.g., blur, white noise, compression, etc.). More specifically, distortion-specific methods can only work when the distortion types are known. In real-world applications, determining the type of distortion itself is a challenging task. Instead, non-distortion-specific algorithms assess the quality of an image without any prior knowledge of distortion types, which brings greater challenges but better applicability. In this paper, we will focus on reviewing non-distortion-specific BIQA methods, and all the BIQA approaches mentioned in the rest of the paper refer to non-distortion-specific methods. In addition, quality assessment for several kinds of images have been studied in previous works, including natural scene (NS) images [13–15], screen content (SC) images [16], depth-image-based-rendering (DIBR)-synthesized images [17,18], 360-degree images [19], etc. Among them, NS images are the most studied, and this review focuses on the blind quality assessment of NS images.

There are a few reviews on the studies of BIQA. Manap et al. [13] presented a survey of non-distortion-specific no-reference IQA, in which they mainly studied the natural-scene-statistics (NSS)-based and learning-based BIQA methods. It should be noted that although this work includes learning-based BIQA methods, it mainly discusses image representation and feature selection based on traditional learning methods such as the codebook approach and principal component analysis (PCA), etc. Xu et al. [14] presented a comprehensive review of BIQA algorithms including both distortion-specific and general-purpose (non-distortion-specific) methods mainly from the perspective of feature extraction and quality prediction. More recently, Yang et al. provided a survey on deep-neural-networks (DNNs)-based BIQA approaches and systematically analyzed these methods according to the role of DNNs [15]. Although there have been previous surveys of BIQA methods, on the one hand, they either focus on traditional algorithms or only cover DNN methods. On the other hand, many new models [20–22] have been proposed since the publication of these reviews. Thus, a new survey is needed to cover both the representative traditional approaches and the very recent advancements in BIQA. To help researchers keep track of the recent progress of BIQA, we aim to present a review of non-distortion-specific BIQA methods covering the most recent advances. The main contributions of this paper are as follows: (1) We presented a formal definition for the BIQA task and proposed a new classification method for BIQA approaches according to the relationship between feature extraction and quality score regression and the method of their realization. (2) We systematically reviewed representative feature extraction techniques and regression models for two-stage methods and typical architectures for end-to-end one-stage approaches, which could help researchers keep track of the recent progress of BIQA. (3) We analyzed the performance of these competing methods and proposed some future research directions based on the discussion of the performance results.

The rest of the paper is organized as follows. Section 2 gives an overview of the BIQA problem and related methods. Section 3 introduces two-stage approaches, and Section 4 presents one-stage approaches. Section 5 presents the commonly used datasets, metrics and analysis of the performance of typical BIQA algorithms. We provide possible future research directions in Section 6 and conclude the paper in Section 7.

2. Overview of Blind Image Quality Assessment

As introduced in the last section, objective BIQA methods aim to assess the quality of an image automatically in the principle of HVS. The mathematical form of the BIQA problem can be described as follows: given an input image X , BIQA methods aim to construct a model ϕ to map the input image X to a scalar quality score or quality rank s as $\phi(X) \rightarrow s$. In most BIQA studies, the BIQA problem is formulated as a regression task as a scalar score in a certain range (e.g., from 0 to 1) that is used to measure the quality of an input image. In addition, in this study, we focus on this kind of BIQA approach.

In order to better introduce the existing works, we grouped the existing BIQA algorithms into two categories (i.e., two-stage and one-stage approaches) based on whether the algorithm handles content representation and quality score prediction separately. The two-stage approaches are the conventional methods, which consist of two independent modules, namely feature extraction and regression. The one-stage approach refers to end-to-end deep neural network models. Figure 1 shows the general workflow for two-stage and one-stage approaches. The two-stage approach contains two separate steps: image content representation (i.e., features) and quality regression. The content representation step aims to extract efficient features for images, while the regression step maps the extracted features to quality scores. Moreover, the features used in two-stage BIQA methods can be further divided into two categories: hand-crafted features and learning-based features. In contrast to two-stage methods, one-stage methods do not explicitly distinguish between content representation and quality regression and, instead, directly map the input image to the quality score.

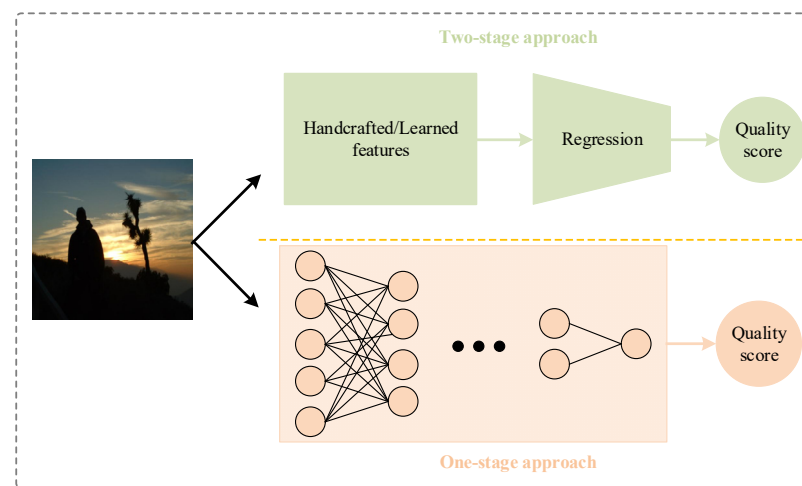


Figure 1. Flowchart of BIQA algorithms. The existing BIQA methods mainly include two categories: two-stage and one-stage approaches.

In two-stage BIQA methods, the input image is first represented using a vector, and this is then used to compute the quality score. Let φ and ψ represent the feature extraction and the quality prediction module, respectively. Then, the two-stage model ϕ can be rewritten as $\psi \circ \varphi$, where the operator \circ indicates using the output of φ as the input for ψ . Note that there may be no model parameters to learn if model φ represents a hand-crafted feature extractor. In fact, most conventional BIQA methods extract various hand-crafted features [8,23–26]. However, with the development of deep learning in recent years, DNN has been used to extract learning-based features for BIQA tasks [27–31], in which the feature extraction module φ becomes φ_{θ_1} with θ_1 as the model parameters that learn the high-level features of an input image X . After acquiring the features (either hand-crafted or learning-based) for an input image, the next step is to compute the quality score by using a trainable regression model ψ (e.g., support vector regression (SVR)). Very recently, end-to-end models have attracted more attention, and many end-to-end DNN models for BIQA have emerged [20,22,31–40]. The end-to-end methods are one-stage approaches as they directly map the input images to the quality scores without an explicit feature extraction. In practice, one-stage BIQA methods utilize DNN models with various architectures to learn a mapping $\phi_{\theta} : X \rightarrow s$ that directly maps an input image X to the quality score s , where θ represents the learnable model parameters. Next, we discuss the technical details of BIQA, including typical hand-crafted and learning-based features, regression models for two-stage BIQA methods, representative DNN architectures for one-stage BIQA approaches, and datasets and metrics for performance evaluation, as well as performance comparison and future research directions.

3. Conventional Two-Stage BIQA Methods

3.1. Hand-Crafted Features

3.1.1. Statistical Features

The natural scene statistics (NSS) model is widely used for reliable hand-crafted feature extraction for IQA. The fundamental idea of NSS is that natural-scene images form a tiny subspace with certain statistical properties (i.e., NSS), and that real-world distortions disturb these statistics [41]. Consequently, features that illustrate the deviation degree of these statistics in distorted images can be used for image quality prediction.

Generalized Gaussian distribution (GGD) is one of the most popular NSS models. Moorthy et al. [23] proposed a method named the blind image quality index (BIQI), in which a wavelet transform over three scales and three orientations is performed on an image. The GGD method is then utilized to model the sub-band coefficients. The GGD is defined as below:

$$f(x; \mu, \sigma^2, \gamma) = a \exp(-[b|x - \mu|]^\gamma) \tag{1}$$

where μ, σ^2 and γ represent the mean, variance and shape parameter of the distribution, respectively, and $a = \frac{b\gamma}{2\Gamma(1/\gamma)}$, $b = (1/\sigma)\sqrt{\frac{\Gamma(3/\gamma)}{\Gamma(1/\gamma)}}$ and $\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt$. For each modelled sub-band, the parameter μ of the fitted GGD is zero, as wavelet bases act as band-pass filters. Consequently, only two parameters (i.e., σ^2 and γ) need to be estimated for each sub-band. Finally, an 18D vector (3 scales \times 3 orientations \times 2 parameters) is extracted as the feature of a distorted image.

Although the features in BIQI are simple and effective, they do not consider the connections between sub-bands. To alleviate this limitation, Moorthy et al. [24] further improved the features in their previous BIQI method by considering the relationship among sub-band coefficients and proposed the distortion-identification-based image verity and integrity evaluation (DIIVINE). They first perform a wavelet transform over two scales and six orientations that results in a 24D (2 scales \times 6 orientations \times 2 parameters) feature vector for a distorted image. In addition, seven features representing the relationships among sub-bands of different scales but with the same orientation are obtained by using GGD fitting. Moreover, 12 features for correlations across scales, 30 features of spatial correlation across sub-bands and 15 features for across-orientation statistics are also extracted. Finally, a total of 88 features are extracted and used for quality prediction, the same as in BIQI.

Mittal et al. [8] proposed the blind/reference-less image spatial quality evaluator (BRISQUE) in the spatial domain. The main idea of BRISQUE is that the normalized luminance of an undistorted image obeys GGD, and the pairwise products of neighboring-normalized luminance in four orientations (i.e., horizontal, vertical, main diagonal and secondary diagonal) obey an asymmetric generalized Gaussian distribution (AGGD). Given an image I with M rows and N columns, the pixel's normalized luminance at (i, j) is calculated as:

$$\tilde{I}(i, j) = \frac{I(i, j) - \mu(i, j)}{\sigma(i, j) + 1} \tag{2}$$

where $\mu(i, j)$ and $\sigma(i, j)$ represent the weighted average and deviation of the three-by-three neighborhood of (i, j) using a 2D circularly symmetric Gaussian weighting function. After GGD is deployed to fit the normalized luminance, the variance and shape parameters of GGD (i.e., σ^2 and γ) form the first two features. In addition, AGGD is utilized to fit the distribution of pairwise coefficients. The definition of AGDD is below:

$$f(x; \sigma_l^2, \sigma_r^2, \gamma) = \begin{cases} \frac{\gamma}{(\beta_l + \beta_r)\Gamma(1/\gamma)} \exp(-\left(\frac{-x}{\beta_l}\right)^\gamma) & x < 0 \\ \frac{\gamma}{(\beta_l + \beta_r)\Gamma(1/\gamma)} \exp(-\left(\frac{-x}{\beta_r}\right)^\gamma) & x \geq 0 \end{cases} \tag{3}$$

where $\beta_l = \sigma_l \sqrt{\Gamma((1/\gamma)/\Gamma(3/\gamma))}$, $\beta_r = \sigma_r \sqrt{\Gamma((1/\gamma)/\Gamma(3/\gamma))}$, σ_l^2, σ_r^2 are the left and right variance, respectively, and γ is the shape parameter. For each paired product, $\sigma_l^2, \sigma_r^2, \gamma$ and the mean of the best AGGD fit together to form the features such that a total of 16 features

are obtained. Finally, a thirty-six-dimensional feature is extracted from two scales and can be further used for quality prediction.

The above methods extract NSS features from the spatial domain, while the frequency domain can provide different perspectives for distortion perception. There are some other NSS-based features using transforms, such as curvelet transform, discrete cosine transform (DCT), etc. Liu et al. [25] proposed the CurveletQA method, which extracts NSS-based features from the curvelet domain, applying the discrete curvelet transform defined below:

$$\theta(j, l, k) = \sum_{0 \leq t_1, t_2 \leq n} f[t_1, t_2] \overline{\varphi_{j,l,k}[t_1, t_2]} \tag{4}$$

where $f[t_1, t_2]$ is a 2D function, $\varphi_{j,l,k}$ represents a curvelet of scale j at position index k with angle index l , and t_1 and t_2 are coordinates in the spatial domain. In the empirical probability distribution function (PDF) $h_j(x) = pdf(\log(|\theta_j|))$, where θ_j is the magnitude of the curvelet coefficients at scale j , $h_j(x)$ is used to effectively capture distribution characters of coefficients with larger amplitude [25]. After that, AGDD, as defined in Formula (3), is used to fit $h_j(x)$. Then, the parameters of the fitted AGDD (i.e., $\sigma_l^2, \sigma_r^2, \gamma$ and the mean of the best AGDD fit), the mean kurtosis, the coefficient of variation in the orientation energy distribution and the energy differences across scales are used to form the 12-D features.

Saad et al. proposed a BLIINDS-II model using NSS-based features from the DCT domain [26]. BLIINDS-II is an improvement in the blind image integrity notator using DCT statistics (BLIINDS) [42]. In BLIINDS, the distorted image is divided into blocks of size 5×5 with a two-pixel overlap, and then a two-dimensional DCT is implemented to compute the local DCT coefficient for each block. The DCT block is further partitioned into sub-regions and sub-bands based on different orientations and frequency bands. More specifically, the DCT block is divided into three oriented sub-regions to take directional information into consideration. In addition, the block is partitioned into low-frequency, mid-frequency and high-frequency sub-bands using the frequency bands strategy. Finally, GGD fitting is applied to each block and to the two partitions (i.e., orientation partition and frequency partition) within the block. The shape parameter γ of the fitted GGD model, the coefficient of frequency variation ζ and the energy sub-band ratio measure R_n are extracted. The frequency variation and energy sub-band ratio measure are defined as follows:

$$\zeta = \sqrt{\frac{\Gamma(1/\gamma)\Gamma(3/\gamma)}{\Gamma^2(2/\gamma)} - 1} \tag{5}$$

$$R_n = \frac{\left| E_n - \frac{1}{n-1} \sum_{j < n} E_j \right|}{E_n + \frac{1}{n-1} \sum_{j < n} E_j} \tag{6}$$

where $E_n = \sigma_n^2$ denotes the average energy in frequency band n , and $n = 1, 2, 3$ correspond to the low-frequency, mid-frequency and high-frequency sub-bands, respectively. A total of 24 features are extracted from three scales.

The summaries mentioned above are representative of hand-crafted features based on NSS models (i.e., GGD and AGGD). Table 1 shows more details about the NSS-based hand-crafted features.

Table 1. Summarization of aforementioned representative NSS-based hand-crafted features.

Method	Transform	NSS Model	Total Features	Description
BIQI [23]	Wavelet transform	GGD	18	Shape parameter and variance of GGD from three orientations over three scales
DIIVINE [24]	Wavelet transform	GGD	88	Improved BIQI by considering the relationship between sub-band coefficients

Table 1. Cont.

Method	Transform	NSS Model	Total Features	Description
BRISQUE [8]	None	GGD and AGGD	36	Model parameters from normalized luminance and pairwise products of neighbouring normalized luminance
CurveletQA [25]	Discrete curvelet transform	AGGD	12	Parameters of AGGD model that fits the logarithm of the magnitude of the curvelet coefficients
BLIINDS-II [26]	DCT	GGD	24	Parameters of GGD model fit for each DCT block and partitions within the block, coefficient of frequency variation and energy sub-band ratio measure, etc.

3.1.2. Texture Features

The texture feature is also used for image quality assessment. In [43], the gradient magnitude (GM) map and the Laplacian of Gaussian (LOG) response are used to capture structural information. The definition of GM map $G(I)$ and LOG response $L(I)$ are as follows:

$$G(I) = \sqrt{(I \otimes h_x)^2 + (I \otimes h_y)^2} \tag{7}$$

$$L(I) = I \otimes h_{LOG} \tag{8}$$

where \otimes denotes the linear convolution operator, h_x is the horizontal Gaussian partial derivative filter, h_y is the vertical Gaussian derivative filter and h_{LOG} is the filter for LOG.

Local binary pattern (LBP) is another texture descriptor that is widely used in various computer vision tasks including image quality assessment. The basic LBP operator takes the following form:

$$LBP_{P,R}(I_c) = \sum_{p=0}^{P-1} S(I_p - I_c)2^p, S(t) = \begin{cases} 1, & \text{if } t \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{9}$$

where P and R stand for the total number of neighbors and the radius of the neighborhood, respectively, I_c is an arbitrary pixel of image I and I_p represents a neighboring pixel of I_c . Figure 2 shows examples of samplings with $R = 2$ and $P = 4, 8, 16$, respectively.

In [44], normalized LBP histograms from different scales are formed as the quality-concerned features. Li et al. [45] proposed a no-reference quality assessment method using statistical structural and luminance features (NRSL). Figure 3 shows the framework of NRSL, in which the distorted image and downscaled images are normalized using Formula (2), and the LBP histogram and luminance histogram are then extracted over multiple scales to form the quality features.

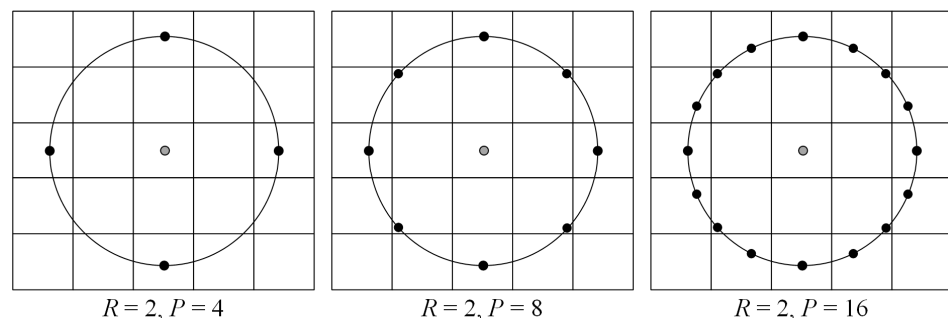


Figure 2. Symmetric samplings with different neighborhood radius R and number of neighbor points P .

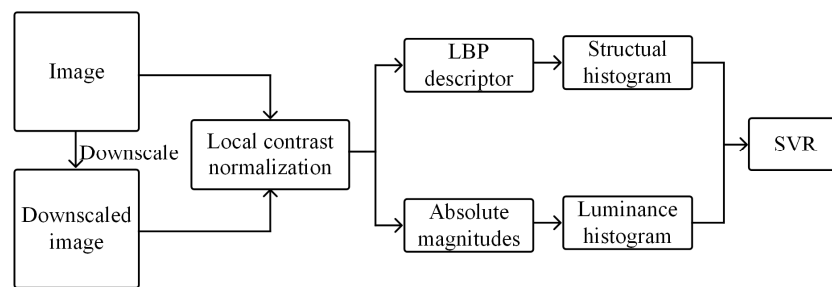


Figure 3. The framework of NRSL model [45].

In [46], Zhang et al. extended the traditional LBP to a generalized LBP (GLBP) as defined below:

$$GLBP_{P,R,T}(I_c) = \sum_{p=0}^{P-1} S(I_p - I_c)2^p, S(t) = \begin{cases} 1, & \text{if } t \geq T \\ 0, & \text{otherwise} \end{cases} \quad (10)$$

where T is a threshold value, and LBP is a special case of GLBP when $T = 0$. In this method, the authors defined a uniformity measure to make GLBP rotation invariant, and the GLBP histograms with W threshold values over D scales are stacked to the final features, where W and D represent the total number of threshold T and scales, respectively.

In addition to GLBP, Freitas et al. have conducted a lot of research to extend LBP descriptors for BIQA, including the local ternary pattern (LTP) method [47], the multiscale local binary patterns (MLBP) method [48], the multiscale salient local binary patterns (MSLBP) method [49] and the orthogonal color planes patterns (OCPP) method [50], etc. In [47], Freitas et al. proposed LTP by extending the coded values in LBP from $\{0, 1\}$ to $\{-1, 0, 1\}$. The step function is defined as per Equation (11) to achieve the development.

$$S(t) = \begin{cases} 1, & t \geq T \\ 0, & -T < t < T \\ -1, & t \leq -T \end{cases} \quad (11)$$

where T is a threshold value that is same as in Equation (10). The ternary pattern is further split into an upper pattern and a lower pattern, corresponding to a positive and negative code, respectively. Figure 4 shows the feature extraction procedure of the LTP descriptor for a single pixel with $R = 1, P = 8$ and $T = 5$. The step function (i.e., Equation (11)) is computed in an order denoted by the numbers in the yellow squares. Different from the LBP descriptor, the LTP descriptor generates three possible values, including $-1, 0$ and 1 , which are represented by the red, black and white colors in Figure 4. The LTP code further splits into two LBP codes for the upper pattern and the lower pattern. For the upper pattern, the negative values (i.e., -1) in the LTP code are converted to 0 . To create the lower pattern, the negative values are converted to 1 , and positive values (i.e., 1) are set to 0 . The two separate LBP channels are then used to calculate the feature vector.

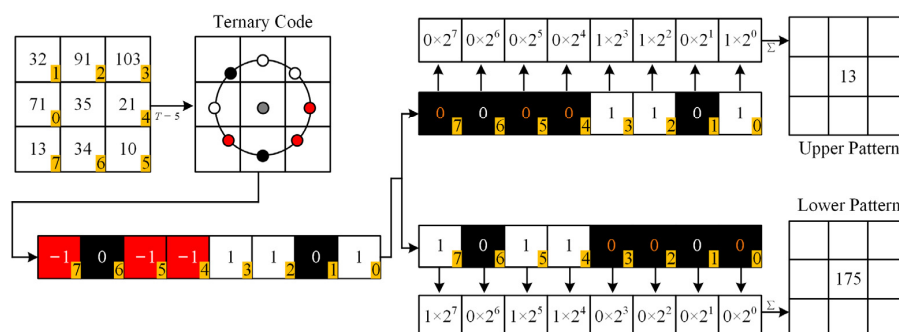


Figure 4. Flowchart for LTP descriptor [47].

MLBP is another representative extension of the LBP proposed by Freitas et al. [48]. The main idea of MLBP is to compute several LBP channels using a different radius of the neighborhood R and number of neighbor pixels, P . MLBP uses the rotation invariant “uniform” pattern descriptor proposed by Ojala et al. [51] instead of the standard LBP described in Equation (9), and the uniform patterns are computed as follows:

$$LBP_{P,R}^u(I_c) = \begin{cases} \sum_{p=0}^{P-1} S(I_p - I_c), & U(LBP_{P,R}^{ri}) \leq 2 \\ P + 1, & otherwise \end{cases}, S(t) = \begin{cases} 1, & \text{if } t \geq 0 \\ 0, & otherwise \end{cases} \quad (12)$$

where $LBP_{P,R}^{ri}(I_c) = \min\{ROTR(LBP_{P,R}(I_c), k), k = 0, \dots, P - 1\}$, $ROTR(x, k)$ denotes the circular bit-wise right shift function shifting x by k positions, and $U(LBP_{P,R}^{ri}) = |S(I_{P-1} - I_c) - S(I_0 - I_c)| + \sum_{p=1}^{P-1} |S(I_p - I_c) - S(I_{p-1} - I_c)|$. It should be noted that the “uniform” LBP reduces the distinct output values from 2^P to $P + 2$ compared to the standard LBP. For a given radius of neighborhood R , there are in total $R + 1$ symmetrical samplings corresponding to $R + 1$ distinct LBP patterns as follows:

$$L_R = \{LBP_{4,R}^u, LBP_{8,R}^u, LBP_{16,R}^u, \dots, LBP_{8R,R}^u\} \quad (13)$$

where $LBP_{P,R}^u$ is computed using Equation (12). If we denote the histogram of each item of L_R as $H_{P,R}$, we can compute it as:

$$H_{P,R} = [h_{P,R}(l_1), \dots, h_{P,R}(l_{P+2})] \quad (14)$$

where $h_{P,R}(l_i) = \sum_{(x,y)} \delta(LBP_{P,R}^u(x, y), i)$, (x, y) denotes the position of the pixel, $\delta(s, t) = 1$ if $s = t$, otherwise $\delta(s, t) = 0$. Then, for a certain radius R , the corresponding histogram can be generated by concatenating all individual LBP histograms as follows:

$$H_R = H_{4,R} \oplus H_{8,R} \oplus \dots \oplus H_{8R,R} \quad (15)$$

where \oplus represents the concatenation operation. Given the maximum radius N , the final feature vector x_N is computed by concatenating all H_R as follows:

$$x_N = H_1 \oplus H_2 \oplus \dots \oplus H_N \quad (16)$$

MSLBP further extends MLBP by introducing visual attention (VA) into the computing of the histograms of $LBP_{P,R}^u$. A saliency map \mathcal{W} is first generated using a VA model (e.g., ITTI [52], TORR [53], etc.), and then the i -th item of the histogram $H_{P,R}$ in Equation (14) is updated as follows:

$$h_{P,R}(l_i) = \sum_{(x,y)} \mathcal{W}(x, y) \delta(LBP_{P,R}^u(x, y), i) \quad (17)$$

Figure 5 presents a diagram of OCPP. The main idea of OCPP is to decompose an image into three orthogonal planes (i.e., XY planes, XZ planes and YZ planes) and extract the LBP features from each individual plane. It should be noted that the sampling of neighboring points in the XZ and YZ planes is different from the standard LBP model as the spatial dimensions vary. For a detailed calculation of the coordinates of the neighboring points in OCPP, please refer to [50,54]. More extended LBP descriptors used in BIQA, including local configuration patterns (LCP) [55], local phase quantization (LPQ) [56], etc., can be found in [54,57].

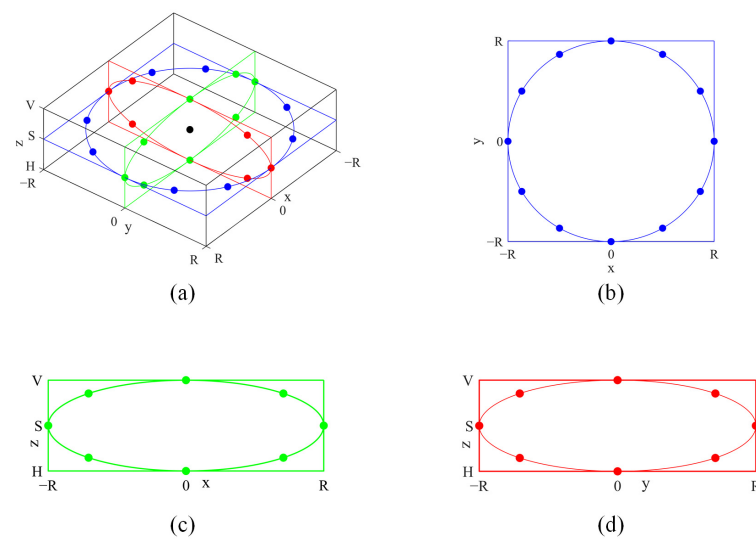


Figure 5. Diagram of OCPP [50]. (a) General view of OCPP, (b) XY plane, (c) XZ plane and (d) YZ plane.

3.1.3. Key Point Descriptors

Besides NSS-model-based features and texture features, key point descriptors (e.g., scale invariant feature transform, SIFT) are also used in BIQA. For instance, Sun et al. [58] proposed a quality feature called SIFT intensity, that is, the total SIFT points in a unit region for quality prediction. In contrast to [58], Nizami et al. [59] proposed the bag-of-features (BOF) method, in which features are constructed using a Harris affine detector, SIFT and K-means. More specifically, the Harris detector is applied to filter image patches containing high-level information reflecting image distortion, from which SIFT points are extracted. The extracted SIFT points are further clustered using the K-means algorithm, and then the cluster centers are used as features for quality assessment. In their implementation, a feature selection step is also utilized to choose optimum features.

3.2. Learning-Based Features

In recent years, deep neural networks (DNN) have been widely used in various computer-vision-related tasks, including BIQA. Although end-to-end is currently the dominant scheme for DNN-based BIQA methods, DNNs were used to learn features in the early days as DNNs can effectively learn high-level semantic features for image representation. Some researchers have proposed learning high-level features from low-level hand-crafted features. Tang et al. [27] extracted middle-level features from NSS, and then they extracted texture and blur features using a deep belief network (DBN). The outputs of the DBN represent the quality features that are further mapped to a quality score using a Gaussian process regression. In [28], the authors constructed a DBN containing three binary–binary-restricted Boltzmann machines (RBMs) to learn NSS features extracted from color spaces and a transform domain. In addition, learning features directly from images or image patches shows promising performance. Li et al. [29] proposed the SFA method, in which a set of overlapping patches are used to represent an image, and an off-the-shelf deep convolutional neural network (DCNN) model is utilized to extract features from these patches. The authors tested three typical pre-trained DCNN models, including AlexNet [60], GoogleNet [61] and ResNet-50 [62], and they chose ResNet-50 as the feature extractor due to its remarkable performance. Sun et al. [30] proposed a BIQA method named GLCP by integrating global high-level semantics and local low-level characteristics. They adopted a DCNN, which was composed of five convolutional layers from a pre-trained AlexNet and one fully connected layer, to extract high-level features. In [31], the authors argued that features extracted from any layer of the DCNN model can be used to represent high-level semantic information. They used each stage of the pre-trained ResNet-50 to form high-level features.

After that, the authors further aggregated the high-level semantics using four different statistical functions to reduce the redundancy of high-dimensional feature data. In [63], Pavan et al. proposed the CONTRastive image quality evaluator (CONTRIQUE), in which an encoder for image representation is trained using a self-supervised contrastive learning strategy. As shown in Figure 6, an input image is processed using several steps including anti-aliasing filtering, down-sampling, random cropping, and color space transforming. Then, the processed image is fed into an encoder (ResNet-50) to generate a representation feature vector. The feature vector is further passed to an MLP predictor, and the output is used to compute the loss value for back-propagation. The infoNCE proposed in [64] is used as a loss function in CONTRIQUE. Once the training is complete, the predictor is discarded, and the outputs of the encoder are used as image representations.

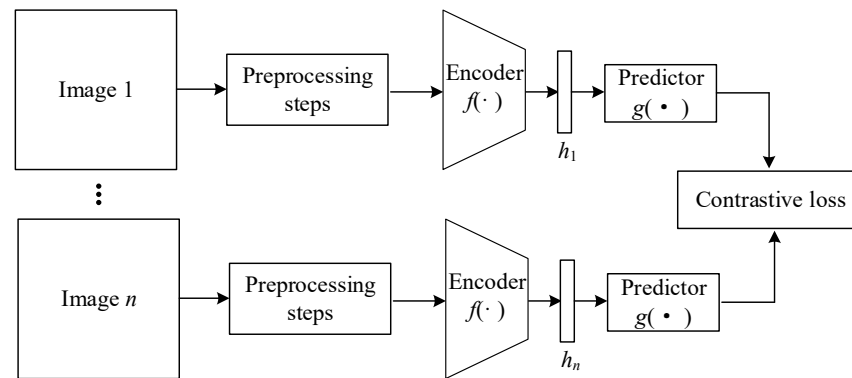


Figure 6. Contrastive learning used in CONTRIQUE [63].

3.3. Quality Regression Models

For two-stage methods, either hand-crafted or learning-based features must be mapped to the quality score. One of the most commonly used regression models is SVR. Given a training dataset $D = \{(x_1, y_1), \dots, (x_l, y_l)\}$, where $x_i \in \mathbb{R}^n$ denotes the quality-aware feature vector and y_i represents the ground truth quality score (i.e., mean opinion score (MOS) or difference mean opinion score (DMOS)), for parameters $C > 0$ and $\epsilon > 0$, SVR can be described as follows [65]:

$$\begin{aligned}
 & \min_{\omega, b, \xi, \xi^*} \frac{1}{2} \omega^T \omega + C \left\{ \sum_{i=1}^l \xi_i + \sum_{i=1}^l \xi_i^* \right\} \\
 & \text{s.t. } \omega^T \phi(x_i) + b - y_i \leq \epsilon + \xi_i \\
 & \quad y_i - \omega^T \phi(x_i) - b \leq \epsilon + \xi_i^* \\
 & \quad \xi_i, \xi_i^* \geq 0, i = 1, \dots, l
 \end{aligned} \tag{18}$$

where $\phi(x_i)$ is a mapping function (i.e., maps x_i to a high-dimensional space), and $K(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ represents a kernel function. For example, the radial basis function (RBF) defined as $K(x_i, x_j) = \exp(-\gamma \|x_i - x_j\|^2)$ is used as a kernel function in [45]. Other studies using SVR include [8,25,30,43,44,46–50,59,66].

In [26,42], a probabilistic model is adopted for quality prediction. More specifically, the multivariate GGD defined in Equation (3) is applied to fit the empirical training data. The distribution fitting $P(X, Y)$ is then used for quality prediction by maximizing $P(y_i|x_i)$.

$$f(x|a, b, \gamma) = a \exp(-(b(x - \mu)^T \Sigma^{-1} (x - \mu))^\gamma) \tag{19}$$

where Σ is the covariance matrix of the multivariate random variable x , and a, b, γ are the same as defined in Equation (1).

In [63], a regularized linear regressor (ridge regression) is used to map the learned representation of images to quality scores. Given a trainable vector $W \in \mathbb{R}^{1 \times n}$, ridge regression aims to learn the optimal W^* using the formula as follows:

$$W^* = \operatorname{argmin}_W \sum_{i=1}^l (y_i - Wx_i)^2 + \lambda \sum_{j=1}^n W_j^2 \tag{20}$$

where x_i, y_i denote the feature vector and ground truth quality score, respectively, as used in SVR, and n denotes the number of dimensions of x_i . In the testing phase, the quality score is computed as $y = Wx$ for an input image represented by feature vector x .

In addition to conventional regression models (i.e., SVR, probabilistic model and ridge regression), DNN models have also been utilized for quality regression. In [67], the authors adopted a general regression neural network (GRNN) for quality regression. Figure 7 shows the architecture of GRNN, which is composed of four layers including the input layer, pattern layer, summation layer and output layer. The number of neurons in the input layer is equal to the dimension of the input feature. The n pattern unit represents n training patterns, where n is the number of training samples. The summation layer contains two units for the assessment of the numerator and denominator of Equation (21), and the output unit computes the quality score based on the two outputs of the summation layer:

$$\bar{y}(x) = \frac{\sum_{i=1}^n y_i \exp(-D_i^2/2\sigma^2)}{\sum_{i=1}^n \exp(-D_i^2/2\sigma^2)} \tag{21}$$

where x is an input feature vector, x_i, y_i represent the feature and quality score of the i -th training sample, respectively, n denotes the training sample number, $D_i^2 = (x - x_i)^T(x - x_i)$ and σ is the spread parameter. Moreover, an MLP, which contains four fully connected layers with the ReLU activation function [68], as shown in Figure 8, is used as a high-capacity regression model in [31] to map high-level semantic features to quality scores. In [69], the authors propose using a DNN model consisting of three hidden layers and one linear regression layer to map extremely large features extracted from a YIQ color space to quality scores. Specifically, each hidden layer is first trained as a sparse auto-encoder to learn the initialization parameters followed by overall fine-tuning.

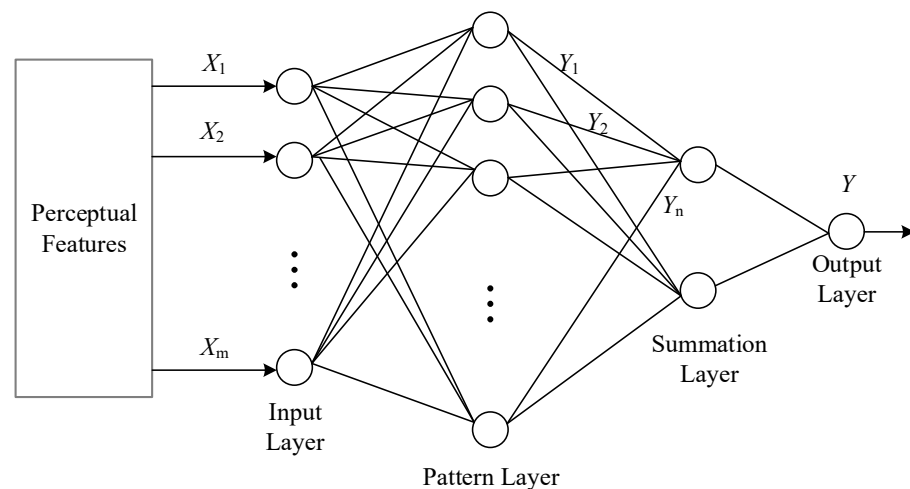


Figure 7. Architecture of GRNN for assessing image quality [67].

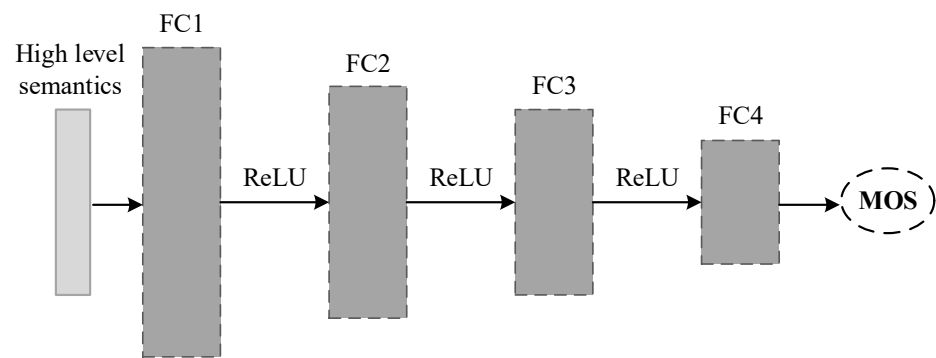


Figure 8. DNN-based regression model for image quality assessment used in [31].

4. DNN-Based One-Stage BIQA Methods

DNN methods have achieved remarkable results in many fields, including BIQA. In this section, we review the representative DNN models for BIQA with various architectures, e.g., the two-stream network, multi-task learning scheme, hyper-network-based model, hierarchical feature fusion, transformer-based method and GAN methods, etc. Next, we will introduce more network details of these representative DNN-based BIQA methods.

4.1. Simple Convolutional Neural Network Models

With the development of deep learning technology and the promising performance of convolutional neural networks (CNN) in various computer vision tasks, researchers have attempted to establish end-to-end BIQA models using CNN. Kang et al. [32] proposed a patch-based CNN network, which was one of the earliest end-to-end DNN models for no-reference image quality assessment. The proposed CNN network is composed of one convolutional layer, two pooling layers (one max pooling and one min pooling), two fully connected layers and a regression layer. More details of each layer are summarized in Table 2. The network input is 32×32 image patches, which are sampled from large images normalized using Formula (2). For the training phase, each patch uses the ground truth score of the original image as its own quality score. The average l_1 norm of the predicted score of each patch and its ground truth score are used to form the loss function, and the model is then optimized using the stochastic gradient descent (SGD) optimizer. In the test stage, the quality score of an image patch is predicted by the trained model, in which the prediction score of a test image is computed by averaging the predicted scores of all test image patches.

Table 2. Detailed configuration of the CNN architecture in [32].

Layer Name	Activation Function	Layer Information
Convolutional layer	/	Fifty kernels with a size of 7×7 and a stride of one pixel
Pooling layer	/	One max pooling and one min pooling
Fully connected layer	ReLU	One fully connected layer with eight hundred neurons
Fully connected layer	ReLU	One fully connected layer with eight hundred neurons
Linear regression layer	/	One fully connected layer with one neuron

Kim et al. [33] proposed a blind image evaluator based on a convolutional neural network (BIECON) to alleviate the accuracy discrepancy between FR-IQA and BIQA. A CNN model $f_{\theta}(\cdot)$ is adopted to extract features from image patches, in which CNN maps a patch to a 100-dimension vector ($\mathbb{R}^{32 \times 32} \rightarrow \mathbb{R}^{100}$). The architecture of the CNN model

is described in Table 3, in which each layer uses ELU [70] as the activation function. A two-step training strategy is utilized to train the model, as shown in Figure 9. In step 1, the divided patches are regressed onto the target local metric scores deriving from a conventional FR-IQA method. Note that a fully connected layer containing only one neuron with a ReLU activation function is added after FC4. In step 2, mean pooling and standard-deviation pooling are adopted to pool features of all patches from a large image. A fully connected layer, which takes the pooled feature vector as input, is used to generate a quality score for the large image. The predicted score and the corresponding ground truth are then used to compute the loss value for model optimization. Similar to [32], the images are also normalized using Formula (2).

Table 3. Detailed configuration of the CNN model in [33].

Layer Name	Activation Function	Layer Information
Conv1	ELU	Forty-eight kernels with a size of 5×5
Max pooling layer	ELU	2×2 max pooling
Conv2	ELU	sixty-four kernels with a size of 5×5
Max pooling layer	ELU	2×2 max pooling
FC1	ELU	One fully connected layer with one thousand six hundred neurons
FC2	ELU	One fully connected layer with four hundred neurons
FC3	ELU	One fully connected layer with two hundred neurons
FC4	ELU	One fully connected layer with one hundred neurons

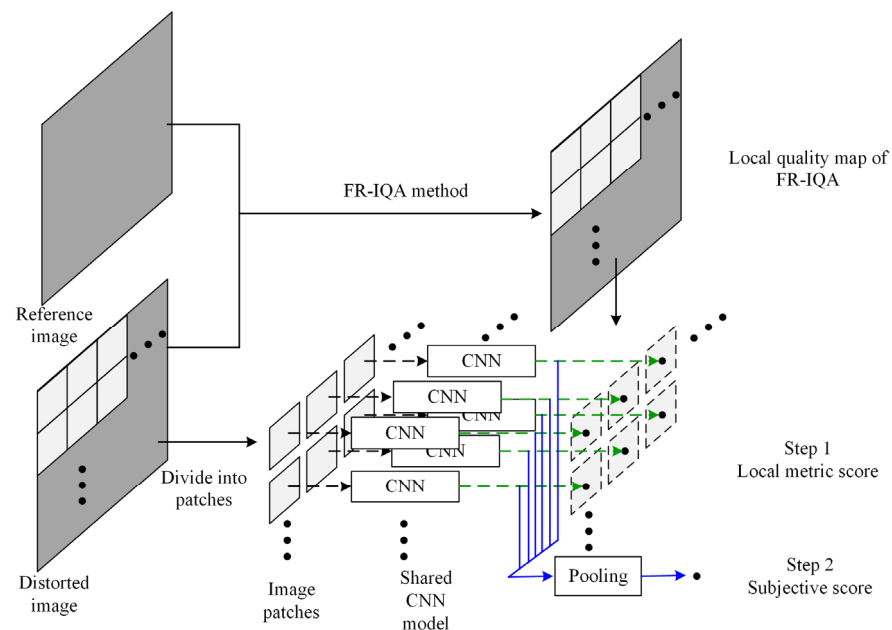


Figure 9. Overall training process of BIECON [33].

As early representative CNN-based models, the above two studies demonstrate the promising prospects of the CNN technique in BIQA even though the models are relatively simple. The main limitation of these earlier networks is that they are not deep enough (consisting of only one or two convolutional layers and several fully connected layers), which greatly affects the models’ feature-learning ability.

4.2. Multi-Task Architectures

Multi-task is another typical scheme used in BIQA. Kang et al. proposed a multi-task CNN model named IQA-CNN++ for simultaneously estimating image quality and identifying distortion [35]. The model takes normalized 32×32 image patches from large images using Formula (2). Figure 10 shows the architecture of IQA-CNN++, which is composed of several shared layers and two task layers. The shared layers include two convolutional layers, three pooling layers and two fully connected layers with ReLU activation functions. Both the linear regression layer and the logistic regression layer take the outputs of the second fully connected layer as inputs for quality prediction and distortion identification, respectively. The loss of the quality prediction is the l_1 norm of the prediction error, and the loss of the distortion classification is a negative log likelihood. Ma et al. [36] also proposed a multi-task model, namely MEON, which takes large images instead of image patches as inputs. Figure 11 shows the model architecture of MEON for two tasks. The shared layers of the two tasks contain four convolutional layers with generalized divisive normalization (GDN) as the activation function and a max pooling layer after each convolutional layer. Two fully connected layers and one softmax layer are appended to predict the distortion type in task 1. Task 2 utilizes two fully connected layers to generate a quality score for each distortion type. A fusion layer (FL) combines the distortion probability generated by task 1 and the perceptual quality scores for each distortion (i.e., the output of the last fully connected layer) to yield an overall quality score. The first fully connected layer of both task 1 and task 2 also use GDN as the activation function.

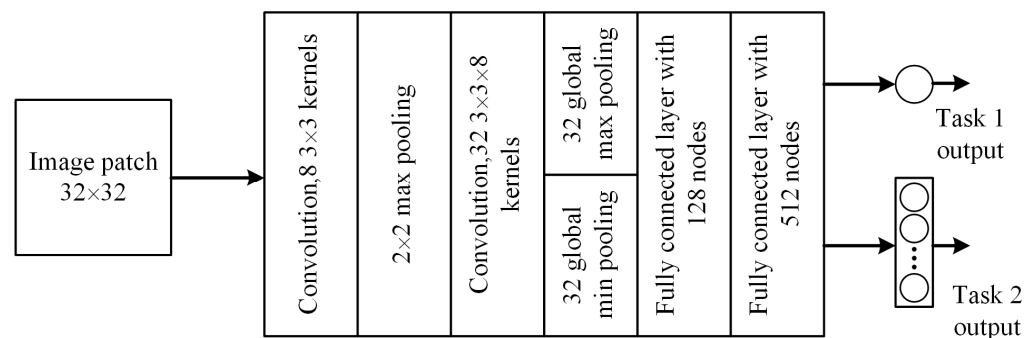


Figure 10. Architecture of the multi-task network IQA-CNN++ [35].

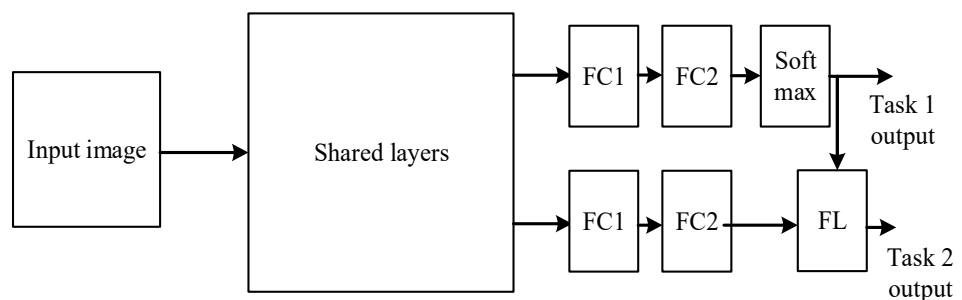


Figure 11. Architecture of the multi-task network MEON [36].

The strength of multi-task architectures is that the network can better perceive the image content through extra tasks, which helps to improve the performance of the model. However, the limitation of the above multi-task models is also obvious. They are only suitable for synthetic distorted images since we cannot know prior information about authentic distortion images, and this limits the application scenes of these models.

4.3. Dual Branch Architectures

To help the model handle more distortion and learn more useful information from different domains, dual branch designs have been used in several BIQA studies.

In [34], the authors proposed a new two-stream convolutional network to learn a more effective feature representation for BIQA from both an RGB image and a gradient image. The architecture of the model is shown in Figure 12, which contains two branches that share the same network structure that contains ten layers. The image stream focuses on capturing information about intensity, and the gradient stream aims to extract structural features from a gradient map. The outputs of the two streams are then concatenated and further fed into a quality regression module to predict the quality score, in which the regression module is composed of two fully connected layers with five hundred and twelve neurons and one linear regression with a one-dimensional output. As shown in Figure 12, both the image stream and the gradient stream take the image patch from large images (i.e., an RGB image and the corresponding gradient map) as the input, and the ground truth score of each large image is assigned to all image patches and gradient patches for model training. In the test phase, the quality score of a large image is obtained by averaging the predicted scores of all patches in the same manner as in [32].

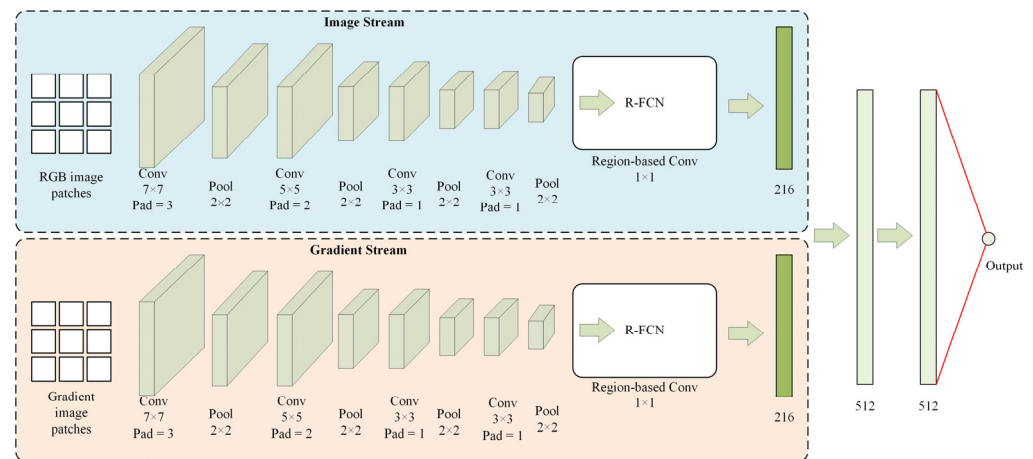


Figure 12. Framework of the two-stream convolutional neural network [34].

To handle both synthetic and authentic distortions, Zhang et al. [37] proposed a deep bilinear CNN (DB-CNN), as shown in Figure 13. The authors designed an S-CNN for synthetic distortions following the style and convention in [71]. Table 4 shows more information about the layers of S-CNN. To train the S-CNN model, a pre-training database containing 852,891 distorted images is constructed using two large-scale databases (i.e., Waterloo Exploration database [72] and PASCAL VOC 2012 [73]). The ground truth is formed as a thirty-nine-dimensional one-hot vector to encode the underlying distortion type at a specific distortion level. Then, the authors tailored the pre-trained S-CNN and VGG-16 (pre-trained on ImageNet) by discarding all layers after the last convolution. A bilinear pooling is appended to fuse the synthetic and authentic features extracted by S-CNN and VGG-16, and finally, a fully connected layer is used to predict the image quality score.

It should be pointed out that for the above dual-branch architectures, the strategy of fusing the features of the two branches is very important. In addition, we should note that in order to improve the representation ability of DB-CNN for synthetic distortion images, the authors constructed a task-related dataset for S-CNN pre-training.

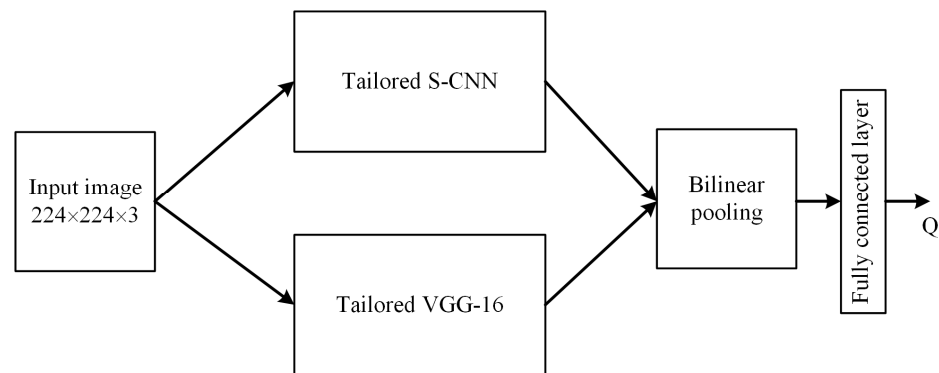


Figure 13. Structure of DB-CNN [37]. The two branches (i.e., tailored S-CNN and tailored VGG-16) are designed for synthetic and authentic features extraction, respectively.

Table 4. Detailed configuration of S-CNN model in [37].

Layer Name	Activation Function	Layer Information
Conv1	ReLU	Forty-eight kernels with a size of 3×3
Conv2	ReLU	Forty-eight kernels with a size of 3×3
Conv3–Conv6	ReLU	Sixty-four kernels with a size of 3×3
Conv7–Conv9	ReLU	One hundred and twenty-eight kernels with a size of 3×3
Average pooling	/	14×14
FC1	ReLU	One fully connected layer with one hundred and twenty-eight neurons
FC2	ReLU	One fully connected layer with two hundred and fifty-six neurons
FC3	ReLU	One fully connected layer with thirty-nine neurons
Softmax	/	Probabilities for thirty-nine classes

4.4. Transformer Based Models

Recently, transformers have attracted a lot of attention [74]. Golestaneh et al. [22] proposed a new architecture based on transformers, relative ranking and self-consistency (TReS) as presented in Figure 14. Similar to [20], the authors also used ResNet-50 to extract the multi-scale features and outputs of all stages, which are concatenated after normalization, pooling and dropout. After that, a transformer encoder following the architecture of [75] and a fully connected layer with ReLU activation are appended, aiming to extract non-local representations of the image. The non-local features and local features (output of the last stage of ResNet-50) are then fused using a fully connected layer to predict the perceptual quality score of the image, as shown in Figure 14. Furthermore, relative ranking loss and self-consistency loss are utilized to consider ranking and correlation between images.

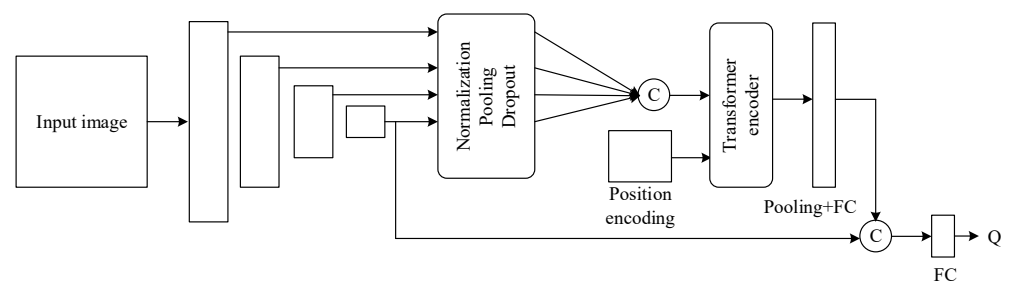


Figure 14. Architecture of TReS [22].

Very recently, Zhu et al. [76] proposed a BIQA method combining self-supervised feature learning and self-attention mechanism for in-the-wild-images. More specifically, they designed a cross-view consistent information mining (CVC-IM) module using the framework of contrastive learning. As shown in Figure 15, the model contains two views, v and u , and two kinds of augmentations based on LAB color images and pseudo-reference images are utilized in contrastive learning to formulate more efficient feature embedding. For the CVC-IM, ResNet50 is employed as the backbone for contrastive learning implementation, and both $L^{v,u}$ and $L^{u,v}$ are implemented using infoNCE loss [64] in the self-supervised learning stage. For the feature-embedding integration, a transformer encoder is employed to implement a self-attention mechanism and map the feature embedding to a quality score. In the following, we refer to this method as CVC-T.

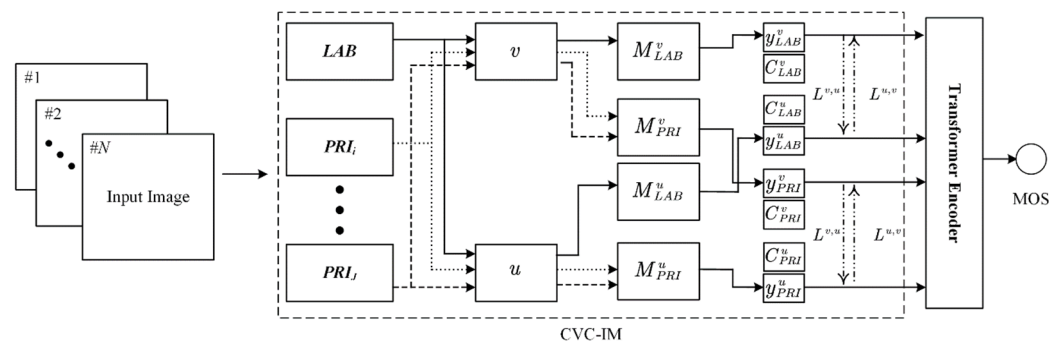


Figure 15. Framework architecture of the CVC-T [76].

Although the above two transformer-based models achieved competitive results on a BIQA task, they also have a significant dependence on data volume, which may indicate that the effective training of a transformer using a small training set is more challenging.

4.5. Other Representative Models

In addition to the above models, there are also many models that design networks from the perspectives of adaptive content perception, multi-scale features and so on.

BIQA for authentically distorted images remains challenging due to the variety of contents and diversity of distortion types. To better deal with images captured in the wild, Su et al. proposed a novel BIQA method named HyperIQA [20] based on a hyper network model [77]. The main idea behind this type of network is to learn how to judge image quality based on the recognition of the image content. The proposed HyperIQA is composed of three parts, including a ResNet-50-based multi-scale feature extraction module, a target network module for image quality prediction and a hyper network module that generates self-adaptive parameters for the target network. More specifically, the multi-scale feature extractor extracts information from the four stages of ResNet-50 and forms the multi-scale features together. The target network contains four fully connected layers that map the multi-scale features to a quality score, in which parameters (i.e., the weights and biases) of all fully connected layers are generated by the content-understanding hyper network. The experimental results showed the superior performance of HyperIQA on wild images and its competitive performance on synthetic distorted images. Figure 16 shows the architecture of HyperIQA. It should be noted that HyperIQA is a model designed for authentic distortion images, and its performance on synthetic images needs further improvement.

To take rich features extracted from CNN into consideration for BIQA, Sun et al. [78] proposed a hierarchical feature fusion (HFF) strategy to hierarchically integrate features from different stages of CNN. Figure 17 presents the architecture of the proposed network, which consists of two parts, a feature-extraction network and a quality regressor. For the feature extraction network, a staircase structure is utilized to fuse the features of different stages of the backbone (i.e., ResNet50). More specifically, a bottleneck structure, which is composed of three convolutional layers, is used to make the channels and dimension

of feature maps of neighboring stages the same. On the other hand, feature maps from different stages are hierarchically merged to avoid the difficulty of network training caused by directly adding the features from lower layers to the final stage. The quality regressor contains three layers, one global average pooling layer and two fully connected layers (with one hundred and twenty-eight neurons and one neuron, respectively). The whole network is then trained in an end-to-end manner using Euclidean distance as the loss function.

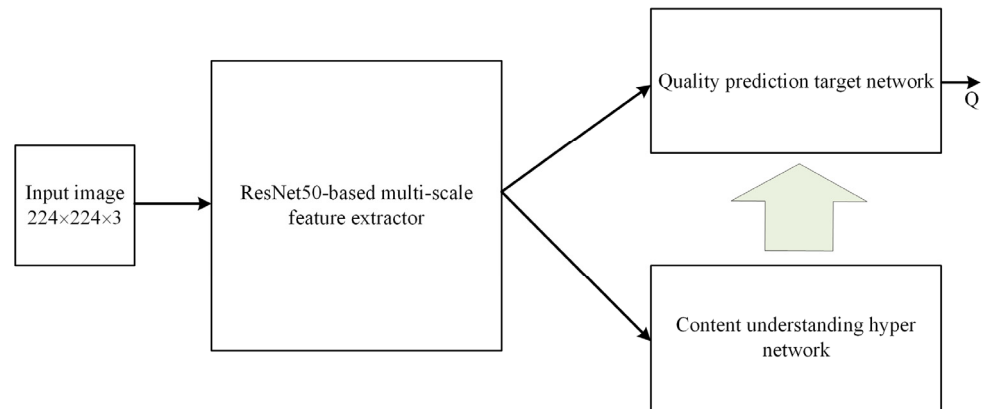


Figure 16. Framework of HyperIQA [20].

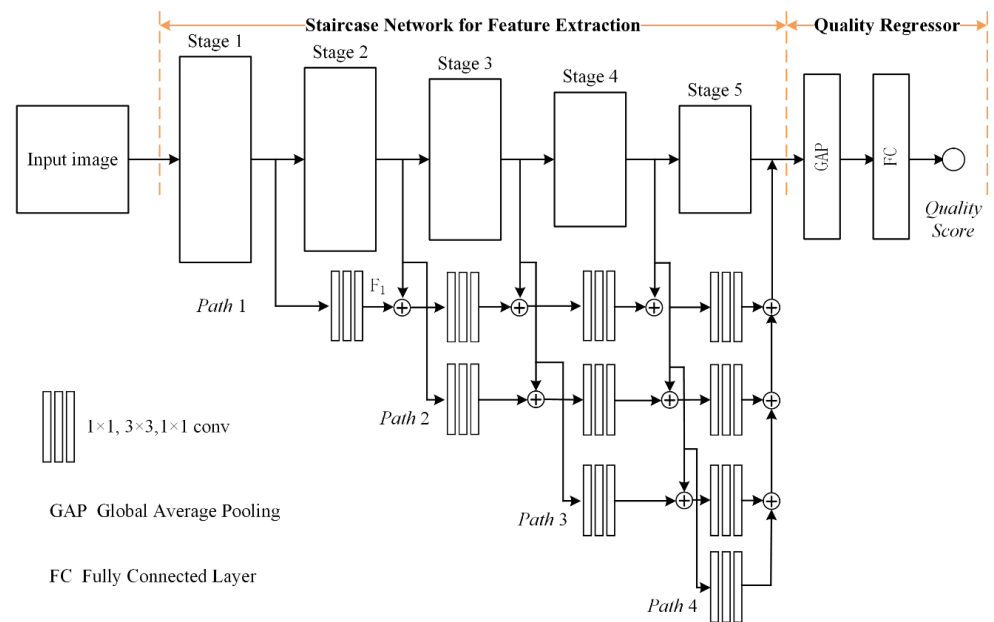


Figure 17. Diagram of the network structure proposed in [78].

Instead of predicting quality scores, Gao et al. [79] formed image quality assessment as an opinion score distribution (OSD) prediction problem, as OSD provides more subjective information than a single MOS. They proposed a CNN model based on fuzzy theory for image OSD prediction, and subsequently, we refer to this method as FOSD-IQA (fuzzy-theory-based OSD IQA). Figure 18 illustrates the image OSD prediction model proposed in [79], and it consists of three modules, namely feature extraction, feature fuzzification and fuzzy transfer. A pre-trained VGG16 discarding all FC layers is used to extract image features, and the output of the last max-pooling layer is flattened into a one-dimensional feature vector. Then, the extracted features are fuzzified using a fuzzy membership function, which is implemented by a convolution operation and an absolute value layer, etc. Finally, the fuzzy feature is mapped to an OSD using an FC layer followed by a Softmax layer.

The Earth mover’s distance (EMD) loss function and a quantile-based loss function are weighted to form the final loss function for end-to-end network training. It should be noted that as the model is designed for OSD prediction, we need to further average the predicted OSD to obtain the predicted MOS.

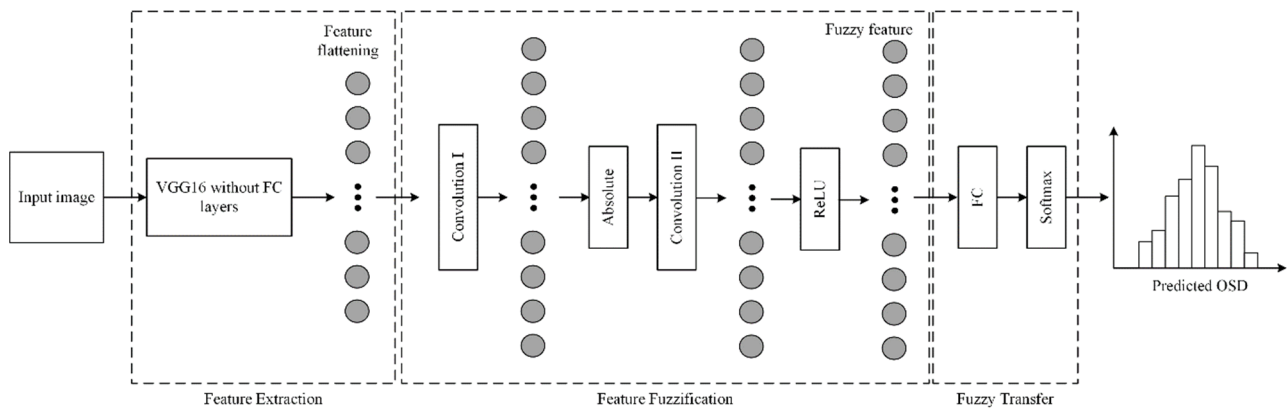


Figure 18. Image OSD prediction model in [79].

In addition, a generative adversarial network (GAN) is also an effective technique for BIQA [38–40]. The main idea of these GAN-based methods is to train a GAN for fake reference image generation, from which quality perception is learned to regress the image quality score.

5. Performance Comparison of BIQA Methods

5.1. Evaluation Metrics

The Spearman rank-order correlation coefficient (SRCC) and Pearson’s linear correlation coefficient (PLCC) are the most commonly used metrics for performance evaluation in BIQA. Both SRCC and PLCC measure the correlation between the predicted quality scores and the subjectively assessed ground true quality scores. For both SRCC and PLCC, a larger value indicates better performance. The definitions of SRCC and PLCC are as follows:

$$SRCC = 1 - \frac{6\sum_i d_i^2}{n(n^2 - 1)} \tag{22}$$

$$PLCC = \frac{\sum_i (q_i - q_m)(\bar{q}_i - \bar{q}_m)}{\sqrt{\sum_i (q_i - q_m)^2 \sum_i (\bar{q}_i - \bar{q}_m)^2}} \tag{23}$$

where d_i represents the rank difference between the predicted quality score and the subjective quality score of the i -th image, and n is the total number of test images. q_i and \bar{q}_i represent the subjective and predicted scores for the i -th test image, respectively, and q_m and \bar{q}_m are the average subjective and predicted quality scores of all images, respectively.

In addition, there are some other commonly used evaluation metrics, such as root mean-squared error (RMSE) and mean absolute error (MAE), which are used to measure the prediction consistency and are defined as per Equations (24) and (25). In the following performance comparison subsection, we only used SRCC and PLCC, as most studies only report SRCC and PLCC results.

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (q_i - \bar{q}_i)^2} \tag{24}$$

$$MAE = \frac{1}{n} \sum_{i=1}^n |q_i - \bar{q}_i| \tag{25}$$

5.2. Typical Datasets

The dataset plays an important role in algorithm evaluation. In this section, we will introduce several typical public synthetic distortion datasets (i.e., LIVE [80], CSIQ [81] and TID2013 [82]) and authentic distortion datasets (i.e., LIVE Challenge (LIVEC) [83], BID [84] and KonIQ-10k [85]) that are widely used in the BIQA field. The LIVE dataset contains twenty-nine reference images and seven hundred and seventy-nine distortion images with five different distortions (i.e., JPEG compression, JP2K compression, white Gaussian noise (WN), gaussian blurring (GB) and fast fading (FF)). The subjective scores of LIVE range from 0 to 100 in the form of a difference mean opinion score (DMOS), with a smaller value indicating a better image quality. The CSIQ dataset is composed of thirty reference images and eight hundred and sixty-six distortion images degraded by six distortions (i.e., JPEG, JP2K, WN, GB, additive pink Gaussian noise (PN) and global contrast decrements (CD)). The subjective scores are also in the form of a DMOS ranging from 0 to 1. Compared with LIVE and CSIQ, the third synthetic distortion dataset TID2013 contains more distortions (i.e., 24 different distortions) and a larger number of images (25 reference images and 3000 distorted images). The subjective score of TID2013 is in the form of a mean opinion score (MOS) ranging from 0 to 9, in which a larger MOS value represents a better image quality. In addition to TID2013, the three authentic datasets LIVEC, BID and KonIQ-10k also use MOS as the subjective scores, in which the MOS values of them are in the range of [0, 100], [0, 5] and [0, 100], respectively. More details of the characteristics of each dataset can be found in Table 5.

Table 5. Details of each IQA dataset. DT stands for distortion type, No. Ref means the number of reference images, No. Dist refers to the number of distorted images, No. DT represents the number of distortion types and SST and RSS mean subjective score type and the range of subjective score, respectively.

Dataset	DT	No. Ref	No. Dist	No. DT	SST	RSS
LIVE	Synthetic	29	779	5	DMOS	[0, 100]
CSIQ	Synthetic	30	866	6	DMOS	[0, 1]
TID2013	Synthetic	25	3000	24	MOS	[0, 9]
LIVEC	Authentic	N/A	1162	-	MOS	[0, 100]
BID	Authentic	N/A	586	-	MOS	[0, 5]
KonIQ-10k	Authentic	N/A	10,073	-	MOS	[0, 100]

5.3. Performance Comparison on Typical IQA Datasets

In this section, we compare the performance of the competing BIQA models on six public datasets. Considering the availability of the source code and the reliability of the results, we directly used the results reported in previous studies.

Table 6 shows the SRCC and PLCC values of each method on three synthetic datasets. For the synthetic distortion datasets (i.e., LIVE, CSIQ and TID2013), deep neural networks, including two-stage methods using learned features and end-to-end one-stage approaches, tended to achieve a better overall performance. However, most conventional hand-crafted feature-based methods (e.g., BRISQUE [8], NRSL [45], NR-GLBP [46]) also obtained competitive results. Compared to the results on LIVE and CSIQ, the results were much worse for all the methods on TID2013, which indicates that the TID2013 is a more challenging dataset. It is worth noting that hand-crafted feature-based results were still promising for synthetic distortions, especially on the datasets with fewer distortion types. For example, BOF-GS [59] obtained the best and second-best results on CSIQ and LIVE, respectively.

Table 7 lists the SRCC and PLCC values of each method on three authentic distortion datasets: LIVEC, BID and KonIQ-10k. Similar to the results on the synthetic datasets, the DNN methods showed great advantages and performed better than the hand-crafted feature-based two-stage approaches. Compared with the results on the synthetic distortion datasets, almost all the hand-crafted feature-based two-stage methods showed a significant

performance degradation on the authentic datasets. Some DNN methods such as DB-CNN, HyperIQA and TReS etc. achieved remarkable performance on both the synthetic and authentic distortion datasets, but the performance of some other DNN models (e.g., GAN-based models, RAN4IQA and CYCLEIQA) also decreased greatly. Furthermore, the size of the datasets had an influence on the performance of the DNN models, as most of the DNN methods on KonIQ-10k performed better than on LIVEC or BID.

Table 6. Performance of each method on synthetic distortion datasets. Two-stage_H and Two-stage_L stand for two-stage method using hand-crafted features and learning-based features, respectively. One-stage denotes one-stage end-to-end deep neural network methods.

Types	Method	Publication Year	LIVE		CSIQ		TID2013	
			SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Two-stage_H	BIQI [23]	2010	0.820	0.821	0.760	0.835	0.349	0.366
	DIIVINE [24]	2011	0.916	0.917	0.835	0.855	0.795	0.794
	BRISQUE [8]	2012	0.940	0.942	0.909	0.937	0.883	0.900
	CurveletQA [25]	2014	0.930	0.933	-	-	-	-
	BLIINDS-II [26]	2012	0.931	0.930	0.900	0.928	0.536	0.538
	Xue's [43]	2014	0.951	0.955	0.924	0.945	-	-
	NR-LBPSriu2 [44]	2013	0.932	0.937	-	-	-	-
	NRSL [45]	2016	0.952	0.956	0.930	0.954	0.945	0.959
	NR-GLBP [46]	2014	0.951	0.954	0.916	0.948	0.920	0.939
	BOF-GS [59]	2020	0.973	0.978	0.971	0.976	0.716	0.718
	LTP [47]	2016	0.942	0.949	0.864	0.880	0.841	-
	MLBP [48]	2016	0.954	-	0.816	-	0.816	-
MSLBP [49]	2018	0.945	-	0.831	-	0.711	-	
OCPP [50]	2018	0.956	-	0.925	-	0.762	-	
Two-stage_L	SFA [29]	2018	0.963	0.972	-	-	0.948	0.954
	GLCP [30]	2016	0.958	0.959	-	-	-	-
	CONTRIQUE [63]	2022	0.969	0.968	0.902	0.927	0.843	0.857
One-stage	CNN [32]	2014	0.956	0.953	-	-	-	-
	BIECON [33]	2016	0.961	0.962	0.815	0.823	0.717	0.762
	Two-stream CNN [34]	2018	0.969	0.978	-	-	-	-
	IQA-CNN++ [35]	2015	0.950	0.950	-	-	-	-
	MEON [36]	2018	0.951	0.955	0.852	0.864	0.808	0.824
	DB-CNN [37]	2018	0.968	0.971	0.946	0.959	0.816	0.865
	HyperIQA [20]	2020	0.962	0.966	0.923	0.942	0.840	0.858
	TReS [22]	2022	0.969	0.968	0.922	0.942	0.863	0.883
	RAN4IQA [38]	2018	0.962	0.967	0.911	0.926	0.816	0.825
	Hall-IQA [39]	2018	0.982	0.982	0.884	0.901	0.879	0.880
CYCLEIQA [40]	2022	0.970	0.971	0.926	0.928	0.832	0.838	

Table 7. Performance of each method on authentic distortion datasets.

Types	Method	Publication Year	LIVEC		BID		KonIQ-10k	
			SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Two-stage_H	BIQI [23]	2010	0.532	0.557	0.573	0.598	-	-
	DIIVINE [24]	2011	0.597	0.627	0.610	0.646	-	-
	BRISQUE [8]	2012	0.607	0.645	0.581	0.605	0.700	0.704
	BLIINDS-II [26]	2012	0.463	0.507	0.532	0.560	0.575	0.584
	NRSL [45]	2016	0.631	0.654	0.638	0.663	-	-
	NR-GLBP [46]	2014	0.612	0.634	0.628	0.654	-	-

Table 7. Cont.

Types	Method	Publication Year	LIVEC		BID		KonIQ-10k	
			SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Two-stage_L	FRIQUEE + DBN [28]	2014	0.672	0.705	-	-	-	-
	SFA [29]	2018	0.812	0.833	0.826	0.840	0.685	0.764
	CONTRIQUE [63]	2022	0.845	0.857	-	-	0.894	0.906
One-stage	CNN [32]	2014	-	-	-	-	0.634	0.671
	BIECON [33]	2016	0.595	0.613	0.539	0.576	0.618	0.651
	MEON [36]	2018	0.697	0.710	-	-	0.611	0.628
	DB-CNN [37]	2018	0.851	0.869	0.845	0.859	0.875	0.884
	HyperIQA [20]	2020	0.859	0.882	0.869	0.878	0.906	0.917
	TReS [22]	2022	0.846	0.877	-	-	0.915	0.928
	RAN4IQA [38]	2018	0.591	0.603	-	-	-	-
	CYCLEIQA [40]	2022	0.786	0.794	-	-	-	-
	HFF [78]	2022	0.862	0.882	0.872	0.883	0.919	0.935
	FOSD-IQA [79]	2022	-	-	-	-	0.905	0.919
	CVC-T [76]	2022	0.872	0.891	-	-	0.915	0.941

6. Discussion and Future Perspectives

Although the existing methods have made great progress in BIQA due to the development of machine learning, especially DNN technologies, BIQA still remains a challenging problem. From the results presented in Section 5.3, we found that most approaches were not suitable for both synthetic and authentic distortions at the same time. In particular, hand-crafted feature-based methods showed great performance degradation on real world authentic distorted images. One intuitive and obvious reason is that the content and distortion types in authentic distorted images are more diverse compared to synthetically distorted images. In fact, the results of the hand-crafted feature-based methods on TID2013, which contains more distortions than LIVE and CSIQ, showed that an increase in the number of distortion types may result in an increase in task difficulty. On the other hand, the results on the authentic distortion datasets showed that the data volume played an important role in the performance of the DNN models. Taking several recently proposed DNN models (e.g., the two-branch method DB-CNN, the hyper-network-based approach HyperIQA and the transformer-based method TReS) as examples, the SRCC values on KonIQ-10k (10,073 real-world images) were greatly improved compared with the results on LIVEC (1162 real-world images). For instance, the SRCC of DB-CNN, HyperIQA and TReS increased by 2.8%, 5.5% and 8.2%, respectively. However, it was difficult to collect sufficient images with labelled quality scores due to the variety of image content, diversity of authentic distortions and the high cost of obtaining image quality scores. Furthermore, although recent DNN models achieved promising performance, the architectures of these models have become more complex, which leads to more limitations on their deployment and application. Based on the analysis above, we summarize some future research directions as follows:

1. Improve the adaptability of DNN models to both synthetic and authentic distortions. It is challenging to adapt to synthetic and authentic distortions at the same time as there are significant differences between them. Although previous efforts have been made to solve this problem (e.g., DB-CNN, see Figure 10), it is an area for further exploration.
2. Build effective BIQA learning models based on limited training samples. The volume of the training set significantly affects the performance of BIQA methods. Since it is difficult to collect sufficient training samples, effective quality assessment is necessary through limited samples.
3. Balance model performance and complexity. Although DNN-based BIQA models have achieved remarkable performance, the models lack deployability due to their model complexity. In fact, simple quality evaluators such as a structure similarity index measure (SSIM) are still widely used due to their simplicity. Thus, the trade-off

between the performance and complexity of DNN-based BIQA models should be taken into account.

7. Conclusions

In this paper, we reviewed BIQA algorithms, including two-stage and one-stage approaches. More specifically, we systematically introduced the representative hand-crafted features, learned features and typical regressors used in two-stage methods and the principle and architecture of various DNN models. We also analyzed the performance of representative BIQA algorithms on six public datasets and suggested future research directions based on the analysis results. This review can provide a helpful reference for researchers interested in the BIQA problem.

Author Contributions: Conceptualization, P.Y., J.S. and L.Q.; methodology, P.Y.; validation, P.Y., J.S. and L.Q.; formal analysis, P.Y.; investigation, P.Y. and J.S.; resources, L.Q.; writing—original draft preparation, P.Y.; writing—review and editing, J.S. and L.Q.; visualization, P.Y.; supervision, L.Q.; funding acquisition, L.Q. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under grant 61866031.

Data Availability Statement: All datasets supporting this study are publicly available.

Conflicts of Interest: The authors declare that we have no conflict of interest in this work.

References

1. Han, J.; Ji, X.; Hu, X.; Zhu, D.; Li, K.; Jiang, X.; Cui, G.; Guo, L.; Liu, T. Representing and retrieving video shots in human-centric brain imaging space. *IEEE Trans. Image Process.* **2013**, *22*, 2723–2736. [[CrossRef](#)] [[PubMed](#)]
2. Tao, D.; Tang, X.; Li, X.; Wu, X. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* **2006**, *28*, 1088–1099. [[PubMed](#)]
3. Tao, D.; Li, X.; Wu, X.; Maybank, S.J. General tensor discriminant analysis and gabor features for gait recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1700–1715. [[CrossRef](#)]
4. Zhu, F.; Shao, L. Weakly-supervised cross-domain dictionary learning for visual recognition. *Int. J. Comput. Vis.* **2014**, *109*, 42–59. [[CrossRef](#)]
5. Li, F.; Shuang, F.; Liu, Z.; Qian, X. A cost-constrained video quality satisfaction study on mobile devices. *IEEE Trans. Multimed.* **2017**, *20*, 1154–1168. [[CrossRef](#)]
6. Wang, Z.; Bovik, A.C.; Sheikh, H.R.; Simoncelli, E.P. Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Process.* **2004**, *13*, 600–612. [[CrossRef](#)]
7. Zhang, L.; Zhang, L.; Mou, X.Q.; Zhang, D. FSIM: A feature similarity index for image quality assessment. *IEEE Trans. Image Process.* **2011**, *20*, 2378–2386. [[CrossRef](#)]
8. Mittal, A.; Moorthy, A.K.; Bovik, A.C. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.* **2012**, *21*, 4695–4708. [[CrossRef](#)]
9. Cheon, M.; Yoon, S.J.; Kang, B.; Lee, J. Perceptual image quality assessment with transformers. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Nashville, TN, USA, 19–25 June 2021; pp. 433–442.
10. Narvekar, N.D.; Karam, L.J. A no-reference image blur metric based on the cumulative probability of blur detection (CPBD). *IEEE Trans. Image Process.* **2011**, *20*, 2678–2683. [[CrossRef](#)]
11. Liu, Y.; Zhai, G.; Gu, K.; Liu, X.; Zhao, D.; Gao, W. Reduced-reference image quality assessment in free-energy principle and sparse representation. *IEEE Trans. Multimed.* **2017**, *20*, 379–391. [[CrossRef](#)]
12. Li, L.; Yan, Y.; Lu, Z.; Wu, J.; Gu, K.; Wang, S. No-reference quality assessment of deblurred images based on natural scene statistics. *IEEE Access* **2017**, *5*, 2163–2171. [[CrossRef](#)]
13. Manap, R.A.; Shao, L. Non-distortion-specific no-reference image quality assessment: A survey. *Inf. Sci.* **2015**, *301*, 141–160. [[CrossRef](#)]
14. Xu, S.; Jiang, S.; Min, W. No-reference/blind image quality assessment: A survey. *IETE Tech. Rev.* **2017**, *34*, 2163–2171. [[CrossRef](#)]
15. Yang, X.H.; Li, F.; Liu, H.T. A survey of DNN methods for blind image quality assessment. *IEEE Access* **2019**, *7*, 123788–123806. [[CrossRef](#)]
16. Gu, K.; Xu, X.; Qiao, J.F.; Jiang, Q.P.; Lin, W.S.; Thalmann, D. Learning a unified blind image quality metric via on-line and off-line big training instances. *IEEE Trans. Big Data* **2019**, *6*, 780–791. [[CrossRef](#)]
17. Yue, G.H.; Hou, C.P.; Gu, K.; Zhou, T.W.; Zhai, G.T. Combining local and global measures for DIBR-synthesized image quality evaluation. *IEEE Trans. Image Process.* **2018**, *28*, 2075–2088. [[CrossRef](#)] [[PubMed](#)]

18. Gu, K.; Qiao, J.F.; Callet, P.L.; Xia, Z.F.; Lin, W.S. Using multiscale analysis for blind quality assessment of DIBR-synthesized images. In Proceedings of the 2017 IEEE International Conference on Image Processing (ICIP), Beijing, China, 17–20 September 2017; pp. 745–749.
19. Sun, W.; Min, X.K.; Zhai, G.T.; Gu, K.; Duan, H.Y.; Ma, S.W. MC360IQA: A multi-channel CNN for blind 360-degree image quality assessment. *IEEE J. Sel. Top. Signal Process.* **2019**, *14*, 64–77. [[CrossRef](#)]
20. Su, S.L.; Yan, Q.S.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; Zhang, Y. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, WA, USA, 13–19 June 2020; pp. 3667–3676.
21. Sun, S.; Yu, T.; Xu, J.; Zhou, W.; Chen, Z. GraphIQA: Learning distortion graph representations for blind image quality assessment. *IEEE Trans. Multimed.* **2022**. [[CrossRef](#)]
22. Golestaneh, S.A.; Dadsetan, S.; Kitani, K.M. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, Waikoloa, HI, USA, 4–8 January 2022; pp. 1220–1230.
23. Moorthy, A.K.; Bovik, A.C. A two-step framework for constructing blind image quality indices. *IEEE Signal Process. Lett.* **2010**, *17*, 513–516. [[CrossRef](#)]
24. Moorthy, A.K.; Bovik, A.C. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Trans. Image Process.* **2011**, *20*, 3350–3364. [[CrossRef](#)]
25. Liu, L.; Dong, H.; Huang, H.; Bovik, A.C. No-reference image quality assessment in curvelet domain. *Signal Process. Image Commun.* **2014**, *29*, 494–505. [[CrossRef](#)]
26. Saas, M.A.; Bovik, A.C.; Charier, C. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Trans. Image Process.* **2012**, *21*, 3339–3352.
27. Tang, H.X.; Joshi, N.; Kapoor, A. Blind image quality assessment using semi-supervised rectifier networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 287–2884.
28. Ghadiyaram, D.; Bovik, A.C. Blind image quality assessment on real distorted images using deep belief nets. In Proceedings of the IEEE Global Conference on Signal and Information Processing, Atlanta, GA, USA, 3–5 December 2014; pp. 946–950.
29. Li, D.Q.; Jiang, T.T.; Lin, W.S.; Jiang, M. Which has better visual quality: The clear blue sky or a blurry animal? *IEEE Trans. Multimed.* **2018**, *21*, 1221–1234. [[CrossRef](#)]
30. Sun, C.R.; Li, H.Q.; Li, W.P. No-reference image quality assessment based on global and local content perception. In Proceedings of the Visual Communications and Image Processing, Chengdu, China, 27–30 November 2016; pp. 1–4.
31. Wang, X.H.; Pang, Y.J.; Ma, X.C. Real distorted images quality assessment based on multi-layer visual perception mechanism and high-level semantics. *Multimed. Tools Appl.* **2020**, *79*, 25905–25920. [[CrossRef](#)]
32. Kang, L.; Ye, P.; Li, Y.; Doermann, D. Convolutional neural networks for no-reference image quality assessment. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, OH, USA, 23–28 June 2014; pp. 1733–1740.
33. Kim, J.; Lee, S. Fully deep blind image quality predictor. *IEEE J. Sel. Top. Signal Process.* **2016**, *11*, 206–220. [[CrossRef](#)]
34. Yan, Q.S.; Gong, D.; Zhang, Y.N. Two-stream convolutional networks for blind image quality assessment. *IEEE Trans. Image Process.* **2018**, *28*, 2200–2211. [[CrossRef](#)]
35. Kang, L.; Ye, P.; Li, Y.; Doermann, D. Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In Proceedings of the IEEE International Conference on Image Processing, Quebec, QC, Canada, 27–30 September 2015; pp. 2791–2795.
36. Ma, K.; Liu, W.; Zhang, K.; Duanmu, Z.; Wang, Z.; Zuo, W. End-to-End Blind Image Quality Assessment Using Deep Neural Networks. *IEEE Trans. Image Process.* **2018**, *27*, 1202–1213. [[CrossRef](#)]
37. Zhang, W.; Ma, K.; Yan, J.; Deng, D.; Wang, Z. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Trans. Circuits Syst. Video Technol.* **2018**, *30*, 36–47. [[CrossRef](#)]
38. Ren, H.Y.; Chen, D.Q.; Wang, Y.Z. RAN4IQA: Restorative adversarial nets for no-reference image quality assessment. In Proceedings of the AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; pp. 7308–7314.
39. Lin, K.Y.; Wang, G.X. Hallucinated-IQA: No-reference image quality assessment via adversarial learning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–23 June 2018; pp. 732–741.
40. Zhang, P.Y.; Shao, X.; Li, Z.H. CycleIQA: Blind Image Quality Assessment Via Cycle-Consistent Adversarial Networks. In Proceedings of the IEEE International Conference on Multimedia and Expo, Taipei, Taiwan, 18–22 June 2022; pp. 1–6.
41. Sheikh, H.R.; Bovik, A.C.; De Veciana, G. An information fidelity criterion for image quality assessment using natural scene statistics. *IEEE Trans. Image Process.* **2005**, *14*, 2117–2128. [[CrossRef](#)]
42. Saad, M.A.; Bovik, A.C.; Charier, C. A DCT statistics-based blind image quality index. *IEEE Signal Process. Lett.* **2010**, *17*, 494–505. [[CrossRef](#)]
43. Xue, W.; Mou, X.; Zhang, L.; Bovik, A.C.; Feng, X. Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features. *IEEE Trans. Image Process.* **2014**, *23*, 4850–4862. [[CrossRef](#)]
44. Zhang, M.; Xie, J.; Zhou, X.; Fujita, H. No reference image quality assessment based on local binary pattern statistics. In Proceedings of the Visual Communications and Image Processing (VCIP), Kuching, Malaysia, 17–20 November 2013; pp. 1–6.
45. Li, Q.H.; Lin, W.S.; Xu, J.T.; Fang, Y. Blind image quality assessment using statistical structural and luminance features. *IEEE Trans. Multimed.* **2016**, *18*, 2457–2469. [[CrossRef](#)]

46. Zhang, M.; Muramatsu, C.; Zhou, X.; Hara, T.; Fujita, H. Blind image quality assessment using the joint statistics of generalized local binary pattern. *IEEE Signal Process. Lett.* **2014**, *22*, 207–210. [[CrossRef](#)]
47. Freitas, P.G.; Akamine, W.Y.L.; Farias, M.C.Q. No-reference image quality assessment based on statistics of local ternary pattern. In Proceedings of the 2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX), Lisbon, Portugal, 6–8 June 2016; pp. 1–6.
48. Freitas, P.G.; Akamine, W.Y.L.; Farias, M.C.Q. Blind image quality assessment using multiscale local binary patterns. *J. Imaging Sci. Technol.* **2017**, *29*, 7–14. [[CrossRef](#)]
49. 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 Proceedings of the 9th ACM Multimedia Systems Conference, Amsterdam, The Netherlands, 12–15 June 2018; pp. 52–63.
50. Freitas, P.G.; Akamine, W.Y.L.; Farias, M.C.Q. No-reference image quality assessment using orthogonal color planes patterns. *IEEE Trans. Multimed.* **2018**, *20*, 3353–3360. [[CrossRef](#)]
51. Ojala, T.; Pietikainen, M.; Maenpaa, T. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Trans. Pattern Anal. Mach. Intell.* **2002**, *24*, 971–987. [[CrossRef](#)]
52. Itti, L.; Koch, C.; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.* **1998**, *20*, 1254–1259. [[CrossRef](#)]
53. Torralba, A.; Oliva, A.; Castelhano, M.S.; Henderson, J.M. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychol. Rev.* **2006**, *113*, 766–786. [[CrossRef](#)]
54. Freitas, P.G.; Da Eira, L.P.; Santos, S.S.; De Farias, M.C.Q. On the application LBP texture descriptors and its variants for no-reference image quality assessment. *J. Imaging* **2018**, *4*, 114. [[CrossRef](#)]
55. Guo, Y.; Zhao, G.; Pietikainen, M. Texture classification using a linear configuration model based descriptor. In Proceedings of the British Machine Vision Conference, Dundee, UK, 29 August–2 September 2011; pp. 119.1–119.10.
56. Ojansivu, V.; Heikkilä, J. Blur insensitive texture classification using local phase quantization. *Lect. Notes Comput. Sci.* **2018**, *5099*, 236–243.
57. Freitas, P.G.; Da Eira, L.P.; Santos, S.S.; Farias, M.C.Q. Image quality assessment using BSIF, CLBP, LCP, and LPQ operators. *Theor. Comput. Sci.* **2020**, *805*, 37–61. [[CrossRef](#)]
58. Sun, T.F.; Ding, S.F.; Xu, X.Z. No-reference image quality assessment through sift intensity. *Appl. Math. Inf. Sci.* **2014**, *8*, 1925–1934. [[CrossRef](#)]
59. Nizami, I.F.; Majid, M.; Rehman, M.U.; Anwar, S.M.; Nasim, A.; Khurshid, K. No-reference image quality assessment using bag-of-features with feature selection. *Multimed. Tools Appl.* **2020**, *79*, 7811–7836. [[CrossRef](#)]
60. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. *Commun. ACM* **2017**, *60*, 84–90. [[CrossRef](#)]
61. Szegedy, C.; Liu, W.; Jia, Y.Q.; 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, Boston, MA, USA, 7–12 June 2015; pp. 1733–1740.
62. He, K.M.; Zhang, X.Y.; Ren, S.P.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 770–778.
63. Madhusudana, P.C.; Birkbeck, N.; Wang, Y.; Adsumilli, B.; Bovik, A.C. Image quality assessment using contrastive learning. *IEEE Trans. Image Process.* **2022**, *31*, 4149–4161. [[CrossRef](#)]
64. Oord, A.V.D.; Li, Y.Z.; Vinyals, O. Representation learning with contrastive predictive coding. *arXiv* **2018**, arXiv:1807.03748.
65. Scholkopf, B.; Smola, A.J.; Bach, F. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*; MIT Press: Cambridge, MA, USA, 2002.
66. Gu, K.; Zhai, G.T.; Yang, X.K.; Zhang, W.J. Using free energy principle for blind image quality assessment. *IEEE Trans. Multimed.* **2014**, *17*, 50–63. [[CrossRef](#)]
67. Li, C.F.; Bovik, A.C.; Wu, X.J. Blind image quality assessment using a general regression neural network. *IEEE Trans. Neural Netw.* **2011**, *22*, 793–799.
68. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the International Conference Machine Learning, Haifa, Israel, 21–24 June 2010; pp. 807–814.
69. Gu, K.; Zhai, G.T.; Yang, X.K.; Zhang, W.J. Deep learning network for blind image quality assessment. In Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), Paris, France, 27–30 October 2014; pp. 511–515.
70. Clevert, D.A.; Unterthiner, T.; Hochreiter, S. Fast and accurate deep network learning by exponential linear units (elus). *arXiv* **2015**, arXiv:1511.07289.
71. Balle, J.; Laparra, V.; Simoncelli, E.P. End-to-end optimized image compression. *arXiv* **2016**, arXiv:1611.01704.
72. 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 Trans. Image Process.* **2016**, *26*, 1004–1016. [[CrossRef](#)]
73. Everingham, M.; Van Gool, L.; Williams, C.K.I.; Winn, J.; Zisserman, A. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* **2010**, *88*, 303–338. [[CrossRef](#)]
74. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Curran Associates: San Francisco, CA, USA, 2017; pp. 5998–6008.

75. Carion, N.; Massa, F.; Synnaeve, G.; Usunier, N.; Kirillov, A.; Zagoruyko, S. End-to-end object detection with transformers. In Proceedings of the European Conference on Computer Vision, Glasgow, UK, 23–28 August 2020; pp. 213–229.
76. Zhu, Y.C.; Li, Y.H.; Sun, W.; Min, X.K.; Zhai, G.T.; Yang, X.K. Blind Image Quality Assessment via Cross-View Consistency. *IEEE Trans. Multimed.* **2022**, 1–14. [[CrossRef](#)]
77. Ha, D.; Dai, A.; Le, Q.V. Hypernetworks. *arXiv* **2016**, arXiv:1609.09106.
78. Sun, W.; Duan, H.Y.; Min, X.K.; Chen, L.; Zhai, G.T. Blind Quality Assessment for in-the-Wild Images via Hierarchical Feature Fusion Strategy. In Proceedings of the 2022 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB), Bilbao, Spain, 15–17 June 2022; pp. 1–6.
79. Gao, Y.X.; Min, X.K.; Zhu, Y.C.; Li, J.; Zhang, X.P.; Zhai, G.T. Image Quality Assessment: From Mean Opinion Score to Opinion Score Distribution. In Proceedings of the 30th ACM International Conference on Multimedia, Lisboa, Portugal, 10–14 October 2022; pp. 997–1005.
80. Sheikh, H.R.; Sabir, M.F.; Bovik, A.C. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.* **2006**, *15*, 3440–3451. [[CrossRef](#)] [[PubMed](#)]
81. Larson, E.C.; Chandler, D.M. Most apparent distortion: Full-reference image quality assessment and the role of strategy. *J. Electron. Imaging* **2010**, *19*, 011006.
82. Ponomarenko, N.; Jin, L.; Ieremeiev, O.; Lukin, V.; Egiazarian, K.; Astola, J.; Vozel, B.; Chehdi, K.; Carli, M.; Battisti, F.; et al. Image database TID2013: Peculiarities, results and perspectives. *Signal Process. Image Commun.* **2015**, *30*, 57–77. [[CrossRef](#)]
83. Ghadivaram, D.; Bovik, A.C. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Trans. Image Process.* **2015**, *25*, 372–387. [[CrossRef](#)]
84. Ciancio, A.; da Costa, A.L.N.T.; da Silva, E.A.B.; Said, A.; Samadani, R.; Obrador, P. No-reference blur assessment of digital pictures based on multifeature classifiers. *IEEE Trans. Image Process.* **2010**, *20*, 64–75. [[CrossRef](#)]
85. Hosu, V.; Lin, H.; Sziranyi, T.; Saupe, D. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Trans. Image Process.* **2020**, *29*, 4041–4056. [[CrossRef](#)]

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.