Research Article Vol. 39, No. 6 / June 2022 / Journal of the Optical Society of America A B1

Journal of the Optical Society of America A

OPTICS, IMAGE SCIENCE, AND VISION

Blind quality assessment of authentically distorted images

LUIGI CELONA* AND RAIMONDO SCHETTINI

Department of Informatics, Systems and Communication, University of Milano-Bicocca, viale Sarca 336, 20126 Milano, Italy
**Corresponding author: luigi.celona@unimib.it*

Received 10 November 2021; revised 12 January 2022; accepted 9 February 2022; posted 10 February 2022; published 2 March 2022

Blind image quality assessment (BIQA) of authentically distorted images is a challenging problem due to the lack of a reference image and the coexistence of blends of distortions with unknown characteristics. In this article, we present a convolutional neural network based BIQA model. It encodes the input image into multi-level features to estimate the perceptual quality score. The proposed model is designed to predict the image quality score but is trained for jointly treating the image quality assessment as a classification, regression, and pairwise ranking problem. Experimental results on three different datasets of authentically distorted images show that the proposed method achieves comparable results with state-of-the-art methods in intra-dataset experiments and is more effective in cross-dataset experiments.

https://doi.org/10.1364/JOSAA.448144

1. INTRODUCTION

In the current day, it is hard to imagine that someone does not have a smartphone in his or her pocket. Some statistics tell us that one person takes 20 photos a day on average in the U.S. These photos are not only taken on special occasions, but they are also taken spontaneously to "seize the moment," without paying particular attention to the light conditions and shooting parameters. The user often takes several photos of the same subject to increase the probability that at least one of them is the "right one," and therefore he or she selects the best shot that he or she will most likely share on social media. But what does it mean that a photo is "right"? According to the International Imaging Industry Association white paper [1], image quality is the "*perceptually weighted combination of all visually significant attributes of an image when considered in its marketplace or application.*" Therefore, we need to consider the application domain and the intended use of an image. The latter, for example, could be used as a visual reference to an item in a digital archive. In this case, it is possible to reasonably assume that the image quality requirements are low, although the image quality is not precisely defined. At this point, we could think that a high-quality image is a faithful reproduction of the original scene or an image that looks the same as the original. However, there are several technical reasons why this type of reproduction is not always possible, but what we have to keep in mind is that, especially in the case of consumer photography, a faithful reproduction of the original is not necessarily the best to pursue.

An observer evaluates a photo with respect to a mental model, which comprises, among others, two dimensions: naturalness and visual aesthetics [2]. *Naturalness* is the degree of correspondence between images and human perception of reality [3], while *visual aesthetics* is a measure of the perceived beauty of a visual stimulus [4]. Both naturalness and visual aesthetics are subjective attributes, but despite this, there are more and more numerous and varied approaches for their automatic estimation starting from collections of subjective data [5,6]. Finally, an image may convey some emotions when it depicts some aspects that are relevant to the persons directly involved with the image, such as the photographer himself, or evokes some pleasant/unpleasant feelings [7].

Having clarified that the estimate of the degree of beauty of images is a very complex and varied problem, in this paper, we concentrate on an unavoidable property: the perceived technical quality of the image. The factors that may influence it are the following:

- Intrinsic to the scene, e.g., geometry and lighting conditions;
- Intrinsic to imaging devices, e.g., spatial resolution, geometric distortions, sharpness, noise, dynamic range, color accuracy, and color gamut;
- Dependent on imaging processing pipelines, e.g., contrast, color balance, color saturation, and compression.

Consumer photographs present an infinite variety of combinations of the above factors, and therefore the automatic assessment of the perceived quality for this type of images is very challenging. Recently, convolutional neural networks (CNNs) have been adopted to estimate the perceived quality of consumer photographs because they can implicitly learn the aforementioned factors in a more effective way than hand-crafted features, whether trained under specific setups [8–10].

In this article, we propose a novel CNN-based blind image quality assessment (BIQA) method. The motivation relies on the fact that BIQA plays an essential role in a broad range of applications, including image acquisition, compression, enhancement, generation, and retrieval [11,12]. Our method may process the entire image to avoid the strong assumption of patch-based BIQA models, namely that the local image quality closely agrees with global subjective scores. However, image quality is inevitably space-varying because of the high degree of nonstationarity of picture contents and the complex perceptual interactions that occur between content and distortions (such as masking). Furthermore, to best mimic the human visual system (HVS), which is sensitive to local distortions when the rest of the image is of good quality [13], we encode the image by combining multi-level features. This design choice is also motivated by the fact that the features of different layers of a CNN network diversely disentangle image distortions [14]. We design and train the proposed method to jointly treat BIQA as a classification, regression, and pairwise ranking problem. There are studies reporting that the use of error metrics (e.g., absolute mean error or mean square error) as optimization criteria have excellent results [15]. However, others show that error metrics achieve poor results compared to cross-entropy, especially for unbalanced datasets or randomly initialized model parameters [16,17]. In particular, some studies have shown the benefits of learning a model for categorizing images into quality levels [18,19]. Therefore, here we propose the so-called score prediction module that estimates both the quality level and the perceived quality score of an image. The quality level is in the form of a discrete level in a scale of values, while the perceived quality score is a continuous value in a predefined range. The parameters of the proposed method for estimating the quality score are optimized end-to-end using a compound loss consisting of a loss for each BIQA problem, namely ordinal cross-entropy for quality-level estimation, mean squared error for quality score regression, and pairwise gap for pairwise quality ranking.

Experimental results on three popular datasets containing authentically distorted images (i.e., LIVE in the Wild Challenge [20], KonIQ-10 k [15], and Smartphone Photography Attribute and Quality [21]) show that the proposed method outperforms state-of-the-art methods. Cross-dataset experiments highlight its high generalization capacity. Finally, the ablation study demonstrates how the combination of BIQA problems and multi-level features contributed to improving the effectiveness of our method.

The remainder of this paper is organized as follows. In Section 2, we review previous methods for image quality assessment and highlight what distinguishes the proposed method from related methods. In Section 3, we describe our method. In Section 4, we provide dataset descriptions, implementation details, and evaluation metrics. In Section 5 we report and analyze results on three datasets for the quality assessment of authentically distorted images, and we conclude in Section 6.

2. RELATED WORK

Predicting the overall quality of an image is the goal of the image quality assessment (IQA) [22,23]. In particular, objective IQA methods are mathematical models capable of predicting a quality score based on human perception, trying to mimic the judgment of the average human observer. The latter's judgment is expressed in terms of the mean opinion score (MOS). Objective IQA can be divided into three groups: full-reference IQA (FR-IQA) methods perform a comparison between the image under test and the reference image [24,25], reduced-reference (RR-IQA) methods use partial information about the reference image [26,27], and no-reference (NR-IQA) methods are used when there is no information on the reference image [15,18,28,29]. In reality, the reference image is usually unavailable. Thus, NR-IQA, also called blind IQA (BIQA), becomes a hot research topic. BIQA methods were first applied to specific types of distortions (e.g., JPEG artifacts and Gaussian blur) using synthetically distorted image databases such as LIVE [30], TID2013 [31], or CSIQ [32]. The above methods do not generalize well and are not suitable for real-world applications. Therefore, in recent years, representative image datasets in terms of authenticity, scale, and diversity have been collected. Datasets such as the LIVE in the Wild Challenge [20] and KonIQ-10 k [15] contain images acquired using consumer cameras, and are therefore possibly affected by a variety of authentic and real-world distortions. Alongside the datasets, powerful new methods have been developed. Considering the applied methodology, BIQA methods can be divided into two categories: hand-crafted feature-based and learning-based.

A. Hand-Crafted Feature-Based BIQA Methods

Many conventional BIQA methods were derived from the natural scene statistics (NSS) model [3]. The main idea behind the NSS model is to measure the distance of an image from the subspace of natural images. The NSS model consists of three steps: (i) extract features from the image, (ii) NSS modeling, and (iii) regression to estimate the overall quality. Based on this framework, Moorthy and Bovik [6] proposed distortion identification-based image verity and integrity evaluation (DIIVINE), which uses steerable wavelet transforms. The features are extracted and classified into the correct distortion type using a support vector machine (SVM) and then a regressor evaluates the image quality. Zhang *et al*. [33] deployed a complex extension of the DIIVINE (C-DIIVINE), which uses a complex steerable pyramid decomposition. Saad *et al*. [34] proposed the BLIINDS index, using the discrete cosine transform (DCT) and later the BLIINDS-II index [35] that uses a Bayesian model to predict the image quality by exploiting the DCT extracted features. Another method that operates in the spatial domain is the blind/referenceless image spatial quality evaluator (BRISQUE), proposed by Mittal *et al*. [36]. The integrated local natural image quality evaluator (IL-NIQE) [37] extends NIQE [38]. It is an opinion-unaware BIQA method that extracts five types of NSS features from pristine images, and uses them to learn a multi-variate Gaussian (MVG) model. The latter then serves as a reference model against which to predict the quality of the test image. Recently, Varga [39] proposed a NR-IQA method in which the image is encoded into a 132-dimensional vector of quality-aware features. A Gaussian process regressor (GPR) is then used to map the previous feature vector into a quality score.

B. Learning-Based BIQA Methods

The blind image quality index (BIQI) [40] follows a two-stage framework in which 25 image quality measures of different distortion-specific features are combined to generate one global quality score. In Chetouani *et al*. [41], the features extracted from three NR metrics (which respectively measure the effect of blocking, blurring, and ringing) are processed by a multilayer perceptron (MLP) for determining the quality score. Ye *et al*. [42] introduced CORNIA, a codebook-based method, that relies on the idea of clustering image patches to create an unlabeled codebook. Then a histogram for quality assessment is obtained for each image by softly assigning patches to the dictionary using pooling strategies. A similar approach is proposed by Xu *et al*. [43]. HOSA uses normalized image patches as local features and soft-assigns each feature to several nearest clusters. Then it uses order statistics between features and clusters in order to get the global quality representation. In recent years, deep learning started to outperform other machine learning techniques in many different fields and applications [44]. Since that moment, a growing number of deep learning BIQA methods have been proposed. Deep learning-based BIQA methods exploit CNN features instead of representing images with handcrafted features. Mainly, these methods rely on backbones pre-trained on big classification datasets (e.g., Imagenet [45]) and extract multiple patches from a single input image. BIECON [46], proposed by Kim and Lee, adopts these described traits: it uses a CNN backbone to extract features and performs a regression onto the local metric score (obtained using a FR-IQA method). In the second step, features pooled from patches are mapped into the subjective quality score of the image. Bianco *et al*. [18] proposed DeepBIQ, which uses multiple image patches and computes the quality score for the whole image by average pooling all the patch scores. Gao *et al*. [47] proposed a similar solution called BLINDER that exploits multi-level representations of images instead of patches. KonCept512 in [15] processes the full-size image and computes the quality score using a pre-trained CNN backbone followed by a global average pooling (GAP). Knowing that saliency is strictly correlated with image quality [48], SGDNet is proposed in [49]. Other BIQA methods rely on the same idea, such as WaDIQaM [50] and VIDGIQA [51]. Many other approaches rely on the *learning to rank* framework: since collecting MOS is very expensive, it is possible to learn quality-aware representations by ranking images with different levels of distortions. RankNet [52], RankIQA [53], and the Siamese networks in [54,55] use this learning paradigm. HyperIQA [56] is a hyper network architecture that exploits the semantic features extracted from a ResNet50 to generate the weights of a quality prediction target network. In addition to the previous features, the ResNet50 extracts multi-level content features that capture both local and global image distortions. These features are input to the quality prediction target network for quality score estimation. Li *et al*. [57] designed Norm-in-Norm, a loss that is closely related to the Pearson linear correlation coefficient (PLCC). It ensures rapid convergence and high effectiveness of the IQA model. Varga [58] presented a novel NR-IQA method that processes the image at three different scales to improve the effectiveness of features extracted from a CNN. The features extracted for each scale are subsequently mapped into a quality score with the help of a GPR. Finally, the quality scores for each scale are averaged.

3. PROPOSED METHOD

The designed architecture for the quality assessment of authentically distorted images is displayed in Fig. 1. Given an input image of any size, the multi-level feature extractor encodes the image into multi-level feature maps. Then, the feature refinement network further processes the previous multi-level features exploiting dilated convolution layers [59] by filtering and aggregating. Finally, the quality score prediction module estimates an image quality level and then maps it into a perceived image quality score. In the following subsections we detail each of the blocks mentioned above motivating the design choices. The contribution of each module is experimentally validated in the ablation study.

A. Multi-Level Feature Extractor

Image quality is commonly measured in terms of MOS for the entire image. However, the perceived image quality varies spatially depending on both the local image content and the perceptual interactions that occur between the content and distortions [13]. Trying to model the previous aspects, the proposed multi-level feature extractor is a CNN backbone that encodes images of varying sizes into multi-level features. In this way we are able to obtain from local- to global-distortion aware features [14]. To this end, we use the Inception-ResNet-v2 [60] model (see Fig. 2) as the backbone. In our network, the last two layers of the original Inception-ResNet-v2, i.e., the average pooling layer and the fully connected layer, are removed. We extract multi-level features from the *Inception-ResNet-A*, *Inception-ResNet-B*, and *Inception-ResNet-C* layers depicted in Fig. 2. Since the feature maps extracted from different layers have different spatial resolutions, we use bilinear downsampling and upsampling to bring them to the same target resolution. Specifically, given an input image with shape $h \times w \times 3$, the feature map for *Inception-ResNet-A* has shape $\frac{h}{8} \times \frac{w}{8} \times 320$, the one for *Inception-ResNet-B* has dimensions $\frac{h}{16} \times \frac{w}{16} \times 1088$, and the one for *Inception-ResNet-C* is $\frac{h}{32} \times \frac{w}{32} \times 1536$, respectively. After resampling, the three feature maps are concatenated along the channel dimension, and the resulting feature map with shape $\frac{h}{16} \times \frac{w}{16} \times 2944$ is input to the feature refinement network as shown in Fig. 1.

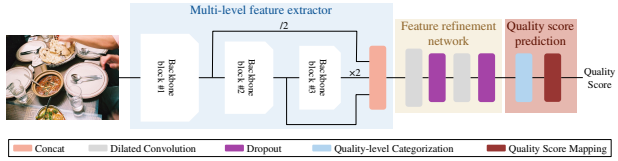


Fig. 1. Architecture of the proposed BIQA method. Given an image, we first extract multi-level features from a pre-trained backbone, and then we feed the feature refinement network that filters and aggregates the previous features. Finally, the score prediction module estimates the perceptual quality score. Batch normalization and ReLU layers after the dilated convolution layers are omitted for simplicity.

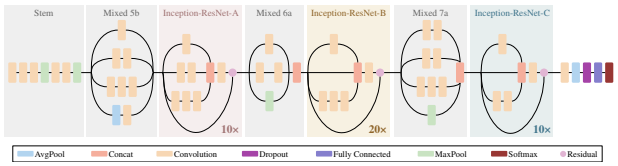


Fig. 2. Inception-ResNet-v2 architecture [60]. Inception-ResNet-A, Inception-ResNet-B, and Inception-ResNet-C are repeated residual blocks.

B. Feature Refinement Network

The obtained feature map is able to capture impairments at different scales. To estimate the perceived quality for the whole image, we refine the previous information and introduce more spatial context by augmenting the network's receptive field using two dilated convolution [59] layers. The previous layers effectively enlarge the receptive field size to incorporate context without introducing extra parameters or computation cost. Thus, the feature map is processed by two layers of dilated convolution with the same dilation term of 2 and kernel of shape 3×3. The first dilated convolution with 1024 output units determines a feature map of $\lfloor \frac{h}{42} \rfloor \times \lfloor \frac{w}{42} \rfloor \times 1024$, while the second one has 512 output units and reduces the feature map to $\frac{h}{64} \times \frac{w}{64} \times 512$. Each dilated convolution is followed by a batch normalization layer, a ReLU activation function, and finally a dropout with a dropout probability of 50% to reduce overfitting. The final quality score prediction module takes as input the feature map having shape $\frac{h}{64} \times \frac{w}{64} \times 512$ and is meant to predict the quality score. It will be detailed in the next section.

C. Quality Score Prediction

The quality score prediction module has a block that categorizes the image into a quality level followed by a second block that maps the discrete category into a continuous quality score.

Quality-level categorization. Some previous methods that rank the quality of an image with respect to K quality anchors have proven effective, mainly due to the benefits of learning using the cross-entropy loss [18,19]. There are many ways to define the anchors. The simplest is to divide the numerical range of subjective scores into a small number of equally spaced intervals. For example, the score range can be partitioned into the five levels of the absolute category ratings (ACR) representing "bad," "poor," "fair," "good," and "excellent" [61]. Or it is also possible to partition the score range, typically [0,1] or [0,100], into K equal bins.

In this work, we express the MOS value, $q_n \in \mathbb{R}$, of a given image n as a quality anchor by rounding the original value, $\bar{q}_n = \lfloor q_n \rceil$. Therefore, $\bar{q}_n \Rightarrow \{x \in \mathbb{Z} : 0 \leq x \leq K\}$ represents the quality level of the image n. The quality-level categorization layer, which consists of a convolution layer with kernel 1×1, outputs the map of logits $\mathbf{Z} = [\mathbf{Z}_0, \mathbf{Z}_1, \ldots, \mathbf{Z}_K]$ for the K quality anchors. Each region of the map is then transformed into a probability distribution thanks to the Softmax layer:

$$\mathbf{p}_i = \frac{\exp(\mathbf{Z}_i)}{\sum_j \exp(\mathbf{Z}_j)}. \tag{1}$$

Finally, a GAP layer is adopted to reduce the map to a probability vector, $\mathbf{p}$. Therefore, we have a quality-level estimate for the entire image.

Quality score mapping. Since existing BIQA methods generally estimate a continuous quality score, we map the quality level to a scalar value for comparison. Given the probability distribution over the K quality anchors, $\mathbf{p} = [p_0, p_1, \ldots, p_K]$, the quality score is obtained as

$$s = \sum_{i=0}^{K} i * p_i. \tag{2}$$

Since $\sum_{i=0}^{K} p_i = 1$, the resulting quality score s is ensured to lie in the range $[0, K]$.

D. Loss Function

The proposed model is end-to-end trained to solve three problems. Indeed, we treat IQA as a classification problem on different quality levels, a regression problem to the quality score, and, finally, a pairwise ranking problem. To this end, we combine the three loss functions that are described below.

Ordinal cross-entropy. The cross-entropy loss is widely used as training loss in classification problems with C categories. Let us denote the input image as $\mathbf{X}$, the ground-truth label vector as $\mathbf{y}$, and the predicted probability distribution as $\mathbf{p}$ of length C. The cross-entropy loss is represented as $L_{\mathrm{CE}}(\mathbf{p}, \mathbf{y}) = -\sum_i^C y_i \log(p_i)$: it focuses only on maximizing the predicted probability of the ground-truth class and ignores the relative distance between an incorrectly predicted data sample and its ground-truth label. However, in the case of ordered classes (e.g., aesthetic and quality assessment), the previous behavior is not ideal because it would be advisable that the most significant errors are those in which the expected score deviates most from the ground-truth label.

Therefore, we define the ordinal cross-entropy (OCE) loss as follows:

$$L_{\mathrm{OCE}}(\mathbf{p}, \mathbf{y}) = -(1 + w) \sum_i^C y_i \log(p_i),$$

$$w = |\mathrm{argmax}(\mathbf{y}) - \mathrm{argmax}(\mathbf{p})|. \tag{3}$$

Here, $(1 + w)$ is a weight term that is multiplied with the regular cross-entropy loss. Within w, argmax returns the index of the maximum valued element in the vector and $|\cdot|$ denotes the absolute value. During the training process, $w = 0$ for training samples that are correctly classified, with the OCE loss being the same as the cross-entropy loss. However, the OCE loss will be higher than cross-entropy loss for misclassified samples and the increase in loss is proportional to how far the samples have been misclassified from their ground-truth label locations.

Mean squared error. Let $(\mathbf{X}, y)$ be the training data, where $\mathbf{X}$ is the input image and y is the corresponding MOS. Given the predicted quality scores $\mathbf{s} = (s_1, \ldots, s_N)$ and the MOS values $\mathbf{y} = (y_1, \ldots, y_N)$ for a batch of N samples, following Li *et al.* [57], we first compute the mean and the L_2-norm of the centered values, for each vector,

$$\hat{a} = \frac{1}{N}\sum_{i=1}^{N} s_i, \quad \hat{b} = \left(\sum_{i=1}^{N} |s_i - \hat{a}|^2\right)^{\frac{1}{2}}, \tag{4}$$

$$a = \frac{1}{N}\sum_{i=1}^{N} y_i, \quad b = \left(\sum_{i=1}^{N} |y_i - a|^2\right)^{\frac{1}{2}}. \tag{5}$$

Second, we normalize the predicted quality scores and the MOS values based on their own statistics, $\hat{s}_i = \frac{s_i - \hat{a}}{\hat{b}}$ and $\hat{y}_i = \frac{y_i - a}{b}$. Thus, the normalized predicted quality scores are $\hat{\mathbf{s}} = (\hat{s}_1, \ldots, \hat{s}_N)$ and the normalized MOS values are $\hat{\mathbf{y}} = (\hat{y}_1, \ldots, \hat{y}_N)$, respectively. We finally use the mean squared error (MSE) as a loss function,

$$L_{\mathrm{MSE}}(\hat{\mathbf{s}}, \hat{\mathbf{y}}) = \frac{1}{N}\sum_{i=1}^{N} |\hat{\mathbf{s}} - \hat{\mathbf{y}}|^2, \tag{6}$$

which is differentiable at the origin, thus being able to produce smoother gradients for small errors than the mean absolute error (MAE), and penalizes larger deviations from the ground truth more heavily.

Pairwise gaps. Although the regression loss in Eq. (6) has implicitly modeled the sorting orders of different images, we also use a pairwise ranking loss to model the score gaps between different images explicitly. Given a batch of N samples consisting of the predicted scores $\hat{\mathbf{s}}$ and the corresponding MOS values $\hat{\mathbf{y}}$ normalized as previously described, the pairwise ranking loss is computed as

$$L_{\mathrm{gaps}}(\hat{\mathbf{s}}, \hat{\mathbf{y}}) = \frac{\sum_{i=1}^{N}\sum_{j>i}^{N} \left|(\hat{s}_i - \hat{s}_j) - (\hat{y}_i - \hat{y}_j)\right|^2}{N(N-1)/2}. \tag{7}$$

L_{gaps} forces the absolute value of the predicted score gap between two images to be no less than the gap between the MOSs to model the sorting relations explicitly.

Compound loss. The loss function used for training the proposed BIQA method is the combination of the aforementioned losses and corresponds to

$$L = \alpha L_{\mathrm{OCE}} + \beta L_{\mathrm{MSE}} + \gamma L_{\mathrm{gaps}}, \tag{8}$$

where α, β, and γ are the trade-off weights, and we set $\alpha = 10$ and $\beta = \gamma = 1$ empirically in all experiments.

E. Discussion

Several works in the literature exploit multi-level features for encoding the whole image [57,62] or image patches [56]. These methods extract feature maps from several convolutional blocks of pre-trained networks. Feature aggregation is achieved after narrowing the maps into vectors by GAP. In neural image assessment (NIMA) [63], the ground-truth distribution of human ratings of a given image is expressed as an empirical probability mass function. A fully connected regression head is then exploited to predict the distribution of ratings. In [18,54], the BIQA methods are trained in two phases: in the first, a CNN learns to categorize images in quality levels, and in the second a regression function is trained to map the representation extracted from the CNN into a quality score. There are three major differences between these methods and the one proposed in this article, which are summarized as follows:

- First, the cited methods narrow multi-level feature maps into vectors by GAP before further processing. In the proposed method, the multi-level feature maps are instead aggregated through the use of dilated convolution layers. Since such layers can be learned, the properties of local distortions are better modeled.
- Second, NIMA's regression head is similar to the proposed quality score mapping module. However, the former is optimized using Earth mover's distance (EMD), while the latter exploits OCE. Our choice is motivated by the fact that EMD has proven effective for estimating distributions, while in our case we maximize the probability for the correct class.
- Finally, differently from [18,54], which learn the BIQA method in two phases, the proposed method is optimized by back-propagation in one easy end-to-end training process because it integrates quality categorization and regression. This aspect accelerates training and leads to a better result.

Apart from the previous differences, there are others that concern more technical aspects, for example, the choice of the interpolation method for feature map resampling. These aspects deserve further investigation, which cannot be included in this work.

4. EXPERIMENTS

In this section we detail the datasets considered for our experiments, the implementation details of the proposed method, and the evaluation metrics.

A. Data

We evaluate our method on three datasets containing consumer photographs, therefore possibly affected by mixtures of generic and authentic distortions. These are the LIVE in the Wild Challenge (LitW) [20] dataset, the KonIQ-10 k (KonIQ) [15] dataset, and Smartphone Photography Attribute and Quality (SPAQ) [21].

LitW consists of 1162 colored images with resolution 500×500 pixels captured from different smartphone cameras. Images are evaluated from 8100 unique subjects via an online crowdsourced user study. More than 350,000 ratings were collected; then the MOS (in the range [0,100]) was computed for each image.

KonIQ consists of 10,073 colored images with a resolution of 1024×768 pixels. Through the use of crowdsourcing, images obtained more than 1.2 million quality ratings from 1459 crowd workers. The MOS in the range [0,100] represents the ground truth for each image.

SPAQ is the largest dataset for BIQA currently available with 11,125 photos taken by 66 smartphones. It contains images affected by a wide range of realistic camera distortions, including sensor noise contamination, out-of-focus blurring, motion blurring, contrast reduction, under-exposure, over-exposure, color shift, and a mixture of multiple distortions above. More than 600 subjects were invited to participate in a subjective test

Table 1. Summary of the Datasets Employed in the Experiments[a]

	LitW [20]	KonIQ [15]	SPAQ [21]
Year	2016	2020	2020
N. images	1,162	10,073	11,125
N. cameras	15[b]	N/A	66
Type of cameras	DSLR/DSC/ Smartphone	DSLR/DSC/Smartphone	Smartphone
Resolution	500 × 500	1024 × 768	Various
Subjective study environment	Crowdsourcing	Crowdsourcing	Laboratory
Extra data	N/A	EXIF	EXIF/Scene cat./Image attr.
MOS range	[0,100]	[0,100]	[0,100]

[a]DSLR, digital single-lens reflex camera; DSC, digital still camera; N/A, not available.
[b]LitW dataset provides the number of manufacturers only.

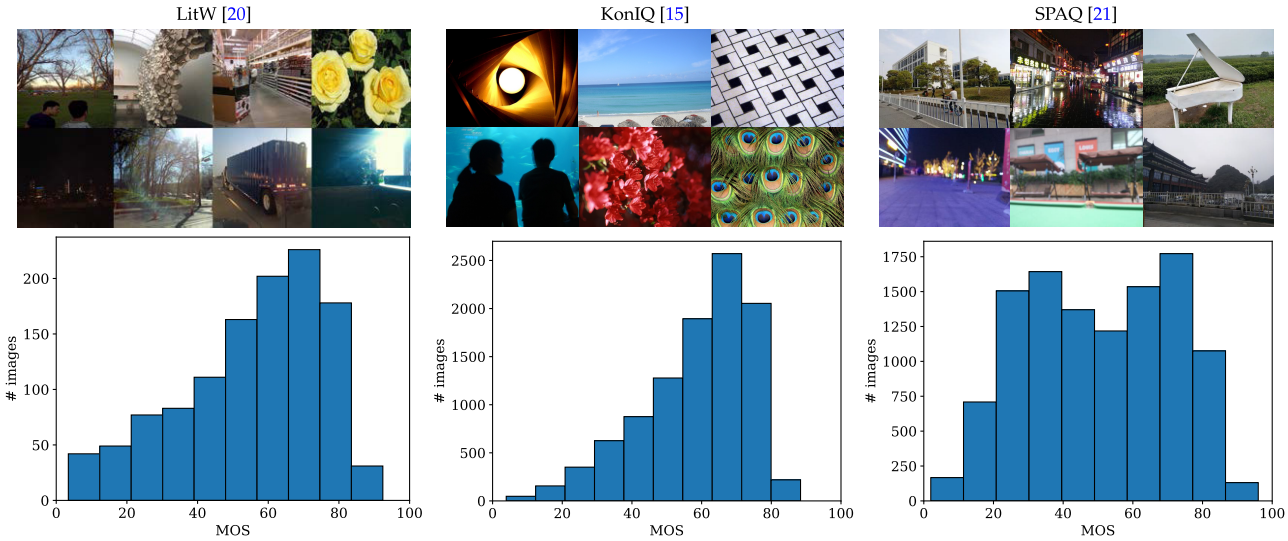


Fig. 3. Sample images and histograms of MOSs of the considered datasets. First row presents sample images from the three considered datasets containing images captured by a wide variety of mobile camera devices and affected by authentic distortions caused by the capture process. The second row shows the histogram of MOS values for the three datasets.

conducted in a well-controlled laboratory setting. The subjects were asked to rate the quality of an image on a continuous scale in [0,100] and provide annotations for five image attributes, namely brightness, hue, contrast, noise, and sharpness, respectively.

A summary of dataset characteristics is provided in Table 1. Figure 3 shows some samples taken from each dataset as well as the histogram of MOS values. As it is possible to see, both LitW and KonIQ are left-skewed MOS distributions. The MOS values of the SPAQ dataset are equally distributed apart from the extremes.

B. Setup

Our method is implemented using PyTorch [64] on a desktop computer with an Intel Core i7-7700 CPU@3.60 GHz, 16 GB DDR4 RAM 2400 MHz, and NVIDIA Titan X Pascal with 3840 CUDA cores. The operating system is Ubuntu 16.04.

Due to the limited number of images that make up current consumer-photo quality datasets, it is impractical to train the InceptionResNet-v2-backbone from scratch. Thus, we exploit the architecture pre-trained to classify the 1.2 million images of ImageNet belonging to 1000 object categories. For all the experiments we use the same setup. Specifically, we exploit the Adam [65] optimizer with an initial learning rate of 10^{-4} and a step scheduler that reduces the learning rate every 15 epochs. The batch size N is equal to 10.

Since the images in the KonIQ and SPAQ datasets have a high resolution that makes them computationally unmanageable, they are resized to 512 × 384 pixels following [15]. During training, we augment the number of images using random horizontal flip and random erasing. Lastly, we select the best model based on the SROCC performance obtained on the validation set.

C. Evaluation Metrics

We evaluate the proposed method using three standard performance metrics: PLCC, Spearman rank order correlation coefficient (SROCC) and residual mean squared error (RMSE). PLCC measures the linear correlation between MOS and predicted scores. It is defined as

$$\mathrm{PLCC} = \frac{\sum_i^N (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i^N (x_i - \bar{x})^2}\sqrt{\sum_i^N (y_i - \bar{y})^2}}, \tag{9}$$

where N is the number of samples, x_i and y_i are the sample at the index i, and $\bar{x}$ and $\bar{y}$ are the means of each sample distribution. We also used SROCC because it is more suitable to represent nonlinear relations. It is defined as follows:

$$\text{SROCC} = 1 - \frac{6\sum_i^N d_i^2}{N(N^2-1)}, \tag{10}$$

with N representing the number of samples and $d_i = (\text{rank}(x_i) - \text{rank}(y_i))$. Lastly, RMSE is used to measure the difference between predicted scores and MOS and is defined as

$$\text{RMSE} = \sqrt{\frac{\sum_i^N \|y_i - s_i\|_2}{N}}, \tag{11}$$

where y_i indicates the MOS, and s_i is the predicted quality score.

5. RESULTS

In this section we report results on the considered datasets. The experimental strategy for each dataset consists of randomly dividing the data into 60% training images, 20% validation images, and the remaining 20% images are used for testing. In order to reduce the influence of random selection, we repeat 10 times the train-val-test split and take the median PLCC, SROCC, and RMSE values as the final result. To ensure a fair comparison with previous methods, we trained and validated state-of-the-art methods on the same data splits. The data splits used for the experiments are made publicly available.

A. Comparison with State-of-the-Art BIQA Methods

To verify the effectiveness of the proposed method compared to previous ones, we include 10 state-of-the-art BIQA methods whose training and evaluation original source codes are available online. We consider both hand-crafted and deep learning-based methods. More in detail, for the first family of methods we have IL-NIQE [37], DIIVINE [6], BRISQUE [36], FRIQUEE [29], CORNIA [42], and HOSA [43]; for the methods based on deep learning instead there are WaDIQaM [50], DeepBIQ [18], HyperIQA [56], KonCept512 [15], and Norm-in-Norm [57]. For each method we used the default settings to run the experiments on our data splits.

Table 2 reports the results in terms of median PLCC, SROCC, and RMSE over 10 random repetitions of train-val-test split. Cells with the symbol "–" indicate that we were unable to run experiments on the adopted hardware with that method for that dataset due to high memory demand issues. From the results obtained it is possible to make various considerations. First, the proposed method outperforms all previous methods on the LitW and SPAQ datasets. On the KonIQ dataset, we obtain lower performance than *Norm-in-Norm* of 0.01 for correlation and 0.7 for RMSE. Second, as was to be expected, deep learning-based BIQA methods far outweigh those based on hand-crafted features. In particular, it could be noted that, although *HOSA* obtained performance very similar to that of *WaDIQaM* on the LitW dataset, its results are significantly lower on the two datasets with higher cardinality and diversity. Third, image-based BIQA methods (such as the one proposed, *Norm-in-Norm*, and *KonCept512*) correlate better with human judgments than patch-based methods (such as *HyperIQA* and *DeepBIQ*). This result confirms that methods that encode the whole image for quality estimation tend to model the interactions that occur between content and distortions more like humans.

Figure 4 shows the scatterplots on the three considered datasets. They report the MOS with respect to the corresponding predicted quality scores for all the samples considered in the 10 iterations. We observe that the method tends to underestimate the quality of images with very high MOS. In SPAQ, the opposite effect also occurs; i.e., for images with very low MOS the predicted quality is overestimated. These behaviors are attributable to the low number of examples at the extremes of the MOS distributions for each dataset.

B. Cross-Dataset Performance Evaluation

In this subsection, we focus on cross-dataset experiments to verify the robustness and generalization capacity of the proposed

Table 2. Comparison with Existing BIQA Methods[a]

	LitW			KonIQ			SPAQ		
Method	**PLCC**	**SROCC**	**RMSE**	**PLCC**	**SROCC**	**RMSE**	**PLCC**	**SROCC**	**RMSE**
IL-NIQE [37]	0.5169	0.4560	39.1210	0.5230*	0.5070*	N/A	0.7210*	0.7130*	N/A
DIIVINE [6]	0.6138	0.5820	16.8672	0.7051	0.6864	11.1381	0.7599	0.7557	14.0174
BRISQUE [36]	0.6402	0.6118	16.3264	0.7001	0.6907	11.2840	0.6966	0.6912	15.1402
FRIQUEE [29]	0.7170	0.6816	14.8289	–	–	–	0.8300*	0.8190*	N/A
CORNIA [42]	0.6759	0.6334	15.7706	0.7950*	0.7800*	N/A	0.7250*	0.7090*	N/A
HOSA [43]	0.6615	0.6285	15.6196	0.7931	0.7664	9.4580	0.7504	0.7452	14.6642
WaDIQaM [50]	0.6746	0.6346	15.3669	0.8050*	0.7970*	N/A	–	–	–
DeepBIQ [18]	0.8512	0.8135	12.6972	0.8962	0.8841	8.9349	–	–	–
HyperIQA [56]	0.8624	0.8321	10.6176	0.9181	0.9032	6.3197	0.9169	0.9135	9.0202
KonCept512 [15]	*0.8789*	0.8484	13.2769	0.9290	0.9102	6.3102	0.9175	0.9141	9.0991
Norm-in-Norm [57]	0.8744	*0.8601*	*10.2356*	**0.9407**	**0.9321**	**5.2477**	*0.9185*	**0.9168**	*8.3088*
Proposed	**0.8819**	**0.8607**	**10.1562**	*0.9346*	*0.9201*	*5.9605*	**0.9192**	*0.9151*	**8.3059**

[a]60% of the images were used for training, 20% for validation, and the remaining 20% for testing. The median PLCC, SROCC, and RMSE over 10 random repetitions are reported for each case. The best and the second-best results on each dataset are marked in bold and italic, respectively. The "-" means that the result cannot be estimated for computational issues. The "*" means that the numbers are taken from [21,56].

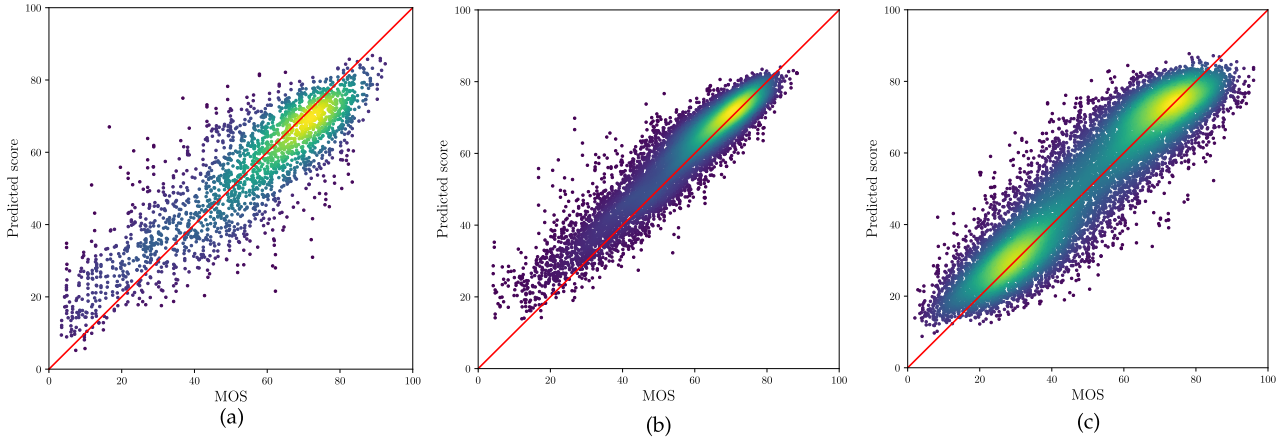


Fig. 4. Scatterplots of the MOS versus the quality score predicted by proposed method versus for (a) LitW test images, (b) KonIQ test images, and (c) SPAQ test images.

Table 3. Cross-Dataset Experiment[a]

Training	LitW		KonIQ		SPAQ	
Testing	KonIQ	SPAQ	LitW	SPAQ	LitW	KonIQ
HyperIQA [56]	0.7282	0.8520	0.7710	0.8313	0.7815	0.7893
KonCept512 [15]	0.7440	0.8571	0.8163	0.8578	**0.7893**	0.7783
Norm-in-Norm [57]	0.7481	0.8596	0.8058	0.8635	0.7679	0.7763
Proposed	**0.7783**	**0.8688**	**0.8180**	**0.8672**	0.7787	**0.8173**

[a]Methods are trained on the "Training" dataset and evaluated on the "Testing" dataset. Performance is reported in terms of SROCC. The best results are marked in bold.

method. In these experiments, the methods are trained on a dataset and tested on another one. More in detail, each method is trained on the whole "training" dataset and evaluated on the entire "testing" dataset. We compare the performance of the proposed method with state-of-the-art methods that achieved similar performance on the three considered datasets, namely *HyperIQA*, *KonCept512*, and *Norm-in-Norm*. The results of the cross-dataset test are summarized in Table 3. It can be seen that our method can outperform the other three NR-IQA methods.

C. Ablation Study

In this subsection we present experiments on the LitW dataset designed to experimentally demonstrate the effectiveness of the design choices.

Combined BIQA problems versus regression. The proposed quality score prediction module and the compound loss for the optimization of the model add complexity to the proposed solution not only in the training phase but also in inference. Thus, we set up an experiment to validate the effectiveness of our proposals. We compare our proposal with a variant in which BIQA is simply treated as a regression problem. In this variant of our method, the quality score prediction module consists only of a linear layer that maps the 512-dimensional features into a perceptual quality score, and the compound loss is replaced with a simple MSE. The results of this solution are reported in terms of median PLCC, SROCC, and RMSE in the row labeled "MSE" of Table 4. As is possible to see from the comparison, the performance drop of the MSE solution is very

Table 4. Comparison between the Proposed Method and Its Variants[a]

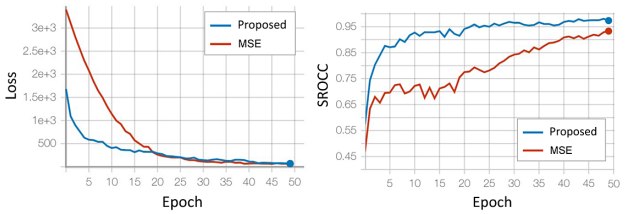
	PLCC	SROCC	RMSE
MSE	0.7684	0.7867	13.9441
Ordinal cross-entropy	0.8532	0.8100	13.7587
Norm. MSE	0.8661	0.8503	12.0853
Pairwise gaps	0.8745	0.8555	11.4514
Single-level features	0.8654	0.8453	11.7324
Proposed	**0.8819**	**0.8607**	**10.1562**

[a]The median PLCC, SROCC, and RMSE over 10 random repetitions are reported for each case. The best results are marked in bold.

significant; in fact the correlations decrease by about 0.08 and the RMSE increases by 3 compared to the *Proposed*.

Another important comparison is provided in Fig. 5. Here, we show loss values and SROCC values across the epochs. We highlight that the *Proposed* has faster convergence than the MSE. The latter in fact, after 50 epochs, has not yet converged and could reach results similar to those of the *Proposed* but at the cost of a greater number of training epochs.

Contribution of each loss function. In this set of experiments, we evaluate the impact on the performance of each of the considered loss functions. Thus, we train the proposed method only using the OCE loss, the MSE loss with normalized inputs, and the pairwise gaps loss. The results are reported in terms of median PLCC, SROCC, and RMSE in Table 4. Among the three losses, the OCE loss achieves the worst results with a PLCC of 0.8532, SROCC equal to 0.8100, and RMSE of 13.7587. The best results are instead achieved by the *pairwise*



(a) (b)

Fig. 5. (a) Loss and (b) SROCC values across epochs on the training set of the LitW dataset.

gaps loss, which obtains a PLCC equal to 0.8745: only 0.01 lower than that of the *Proposed* (i.e., PLCC = 0.8819).

Multi- versus single-level features. In the third experiment, we empirically demonstrate that multi-level features provide better image encoding for estimating its quality. To this end, the input image with shape $h \times w \times 3$ is encoded by the multi-level feature extractor using only the map consisting of $\frac{h}{32} \times \frac{w}{32} \times 1536$ features extracted from the *Inception-ResNet-C* level. This variant is labeled as "single-level features" in Table 4. Using single-level features rather than multi-level features results in minimal loss of effectiveness. In fact, single-level SROCC corresponds to 0.8621, while multi-level SROCC is equal to 0.8522.

6. CONCLUSION

In this paper, a novel architecture for authentically distorted images BIQA has been proposed. Given an input image, our model extracts the features of different layers to better encode image impairments. The multi-level features are then used to estimate a quality level for the input image. Finally, the quality level is mapped into a perceptual quality score. The proposed architecture is trained end-to-end by treating BIQA jointly as a classification, regression, and pairwise ranking problem. Extensive experiments have been carried out on three BIQA datasets with authentic distortions, such as LIVE In the Wild [20], KonIQ-10 k [15], and SPAQ [21]. The introduced method is able to achieve good performance in intra-dataset experiments. The results obtained for the cross-dataset experiments show the high generalization capacity of the proposed method. From the ablation study it emerges that treating BIQA as a compound problem determines a performance gain. The proposed method overestimates images whose MOS are very low and underestimates images whose MOS are very high. From a computational point of view, the method is not suitable for deployment on devices with limited computational resources. In fact, InceptionResNet-v2 used as a backbone to extract multi-level features consists of a large number of parameters and operations. Lightweight CNN architectures might be considered in the future for improving the method efficiency.

To facilitate the reproducibility of the presented results, the source code of the proposed method, the pre-trained models, and the split train-val-tests are available at [66].

Acknowledgment. We thank Francesco Prete for his helpful support of this project during his stage at the Imaging and Vision lab.

Disclosures. The authors declare no conflicts of interest.

Data availability. Data underlying the results presented in this paper are available in Refs. [15,20,21,66].

REFERENCES

1. "Fundamentals and review of considered test methods," CPIQ Initiative Phase 1 White Paper (International Imaging Industry Association, 2007).
2. C. Batini and M. Scannapieco, *Data and Information Quality* (Springer, 2016).
3. A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," Springer J. Math. Imaging Vis. **18**, 17–33 (2003).
4. C. J. Cela-Conde, G. Marty, F. Maestú, T. Ortiz, E. Munar, A. Fernández, M. Roca, J. Rosselló, and F. Quesney, "Activation of the prefrontal cortex in the human visual aesthetic perception," Proc. Natl. Acad. Sci. USA **101**, 6321–6325 (2004).
5. L. Celona and R. Schettini, "A genetic algorithm to combine deep features for the aesthetic assessment of images containing faces," MDPI Sens. **21**, 1307 (2021).
6. A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," IEEE Trans. Image Process. **20**, 3350–3364 (2011).
7. B. Keelan, *Handbook of Image Quality: Characterization and Prediction* (CRC Press, 2002).
8. L. Celona and R. Schettini, "CNN-based image quality assessment of consumer photographs," in *London Imaging Meeting* (Society for Imaging Science and Technology, 2020), Vol. **2020**, pp. 129–133.
9. R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Conference on Computer Vision and Pattern Recognition* (IEEE, 2018), pp. 586–595.
10. X. Yang, F. Li, and H. Liu, "A survey of DNN methods for blind image quality assessment," IEEE Access **7**, 123788 (2019).
11. Z. Wang, "Applications of objective image quality assessment methods [applications corner]," IEEE Signal Process. Mag. **28**(6), 137–142 (2011).
12. K. Ma and Y. Fang, "Image quality assessment in the modern age," in *International Conference on Multimedia* (ACM, 2021), pp. 5664–5666.
13. P. Bex, "Sensitivity to spatial distortion in natural scenes," J. Vision **8**, 688 (2008).
14. S. Bianco, L. Celona, and P. Napoletano, "Disentangling image distortions in deep feature space," Elsevier Pattern Recogn. Lett. **148**, 128–135 (2021).
15. V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: an ecologically valid database for deep learning of blind image quality assessment," IEEE Trans. Image Process. **29**, 4041–4056 (2020).
16. S. Baccianella, A. Esuli, and F. Sebastiani, "Evaluation measures for ordinal regression," in *International Conference on Intelligent Systems Design and Applications* (IEEE, 2009), pp. 283–287.
17. P. Golik, P. Doetsch, and H. Ney, "Cross-entropy vs. squared error training: a theoretical and experimental comparison," in *Interspeech* (2013), Vol. **13**, pp. 1756–1760.
18. S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," Springer Signal, Image Video Process. **12**, 355–362 (2018).
19. H. Zeng, L. Zhang, and A. C. Bovik, "Blind image quality assessment with a probabilistic quality representation," in *International Conference on Image Processing (ICIP)* (IEEE, 2018), pp. 609–613.
20. D. Ghadiyaram and A. C. Bovik, "Massive online crowd sourced study of subjective and objective picture quality," IEEE Trans. Image Process. **25**, 372–387 (2016).
21. Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *CVPR* (IEEE, 2020), pp. 3677–3686.
22. M. Li, W. Zuo, S. Gu, D. Zhao, and D. Zhang, "Learning convolutional networks for content-weighted image compression," in *CVPR* (IEEE, 2018), pp. 3214–3223.

23. W.-S. Lai, J.-B. Huang, Z. Hu, N. Ahuja, and M.-H. Yang, "A comparative study for single image blind deblurring," in *CVPR* (IEEE, 2016), pp. 1701–1709.
24. J. Ma, P. An, L. Shen, and K. Li, "Full-reference quality assessment of stereoscopic images by learning binocular visual properties," Appl. Opt. **56**, 8291–8302 (2017).
25. Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," IEEE Trans. Image Process. **13**, 600–612 (2004).
26. Y. Liu, G. Zhai, K. Gu, X. Liu, D. Zhao, and W. Gao, "Reduced-reference image quality assessment in free-energy principle and sparse representation," IEEE Trans. Multimedia **20**, 379–391 (2017).
27. Z. Wang and E. P. Simoncelli, "Reduced-reference image quality assessment using a wavelet-domain natural image statistic model," Proc. SPIE **5666**, 149–159 (2005).
28. S. Gabarda and G. Cristóbal, "Blind image quality assessment through anisotropy," J. Opt. Soc. Am. A **24**, B42–B51 (2007).
29. D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," J. Vis. **17**(1), 32 (2017).
30. H. R. Sheikh, Z. Wang, L. Cormack, and A. C. Bovik, *Live Image Quality Assessment Database Release 2* (2005).
31. N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Color image database TID2013: peculiarities and preliminary results," in *European Workshop on Visual Information Processing (EUVIP)* (IEEE, 2013), pp. 106–111.
32. E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," J. Electron. Imaging **19**, 011006 (2010).
33. Y. Zhang, A. K. Moorthy, D. M. Chandler, and A. C. Bovik, "C-DIIVINE: no-reference image quality assessment based on local magnitude and phase statistics of natural scenes," Elsevier Signal Process. Image Commun. **29**, 725–747 (2014).
34. M. A. Saad, A. C. Bovik, and C. Charrier, "A DCT statistics-based blind image quality index," IEEE Signal Process. Lett. **17**, 583–586 (2010).
35. M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: a natural scene statistics approach in the DCT domain," IEEE Trans. Image Process. **21**, 3339–3352 (2012).
36. A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," IEEE Trans. Image Process. **21**, 4695–4708 (2012).
37. L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," IEEE Trans. Image Process. **24**, 2579–2591 (2015).
38. A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," IEEE Signal Process. Lett. **20**, 209–212 (2012).
39. D. Varga, "No-reference image quality assessment with global statistical features," MDPI J. Imaging **7**, 29 (2021).
40. A. K. Moorthy and A. C. Bovik, "A two-step framework for constructing blind image quality indices," IEEE Signal Process. Lett. **17**, 513–516 (2010).
41. A. Chetouani, A. Beghdadi, S. Chen, and G. Mostafaoui, "A free reference image quality measure using neural networks," in *International Workshop on Video Processing and Quality Metrics* (2010).
42. P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *CVPR* (IEEE, 2012), pp. 1098–1105.
43. J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," IEEE Trans. Image Process. **25**, 4444–4457 (2016).
44. A. Voulodimos, N. Doulamis, A. Doulamis, and E. Protopapadakis, "Deep learning for computer vision: a brief review," Comput. Intell. Neurosci. **2018**, 7068349 (2018).
45. A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems* (2012), pp. 1097–1105.
46. J. Kim and S. Lee, "Fully deep blind image quality predictor," IEEE J. Sel. Top. Signal Process. **11**, 206–220 (2016).
47. F. Gao, J. Yu, S. Zhu, Q. Huang, and Q. Tian, "Blind image quality prediction by exploiting multi-level deep representations," Elsevier Pattern Recogn. **81**, 432–442 (2018).
48. L. Zhang, Y. Shen, and H. Li, "VSI: a visual saliency-induced index for perceptual image quality assessment," IEEE Trans. Image Process. **23**, 4270–4281 (2014).
49. S. Yang, Q. Jiang, W. Lin, and Y. Wang, "SGDNET: an end-to-end saliency-guided deep neural network for no-reference image quality assessment," in *International Conference on Multimedia* (ACM, 2019), pp. 1383–1391.
50. S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," IEEE Trans. Image Process. **27**, 206–219 (2017).
51. J. Guan, S. Yi, X. Zeng, W.-K. Cham, and X. Wang, "Visual importance and distortion guided deep image quality assessment framework," IEEE Trans. Multimedia **19**, 2505–2520 (2017).
52. C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender, "Learning to rank using gradient descent," in *International Conference on Machine Learning (ICML)* (2005), pp. 89–96.
53. X. Liu, J. van de Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," in *International Conference on Computer Vision (ICCV)* (IEEE, 2017), pp. 1040–1049.
54. W. Zhang, K. Ma, and X. Yang, "Learning to blindly assess image quality in the laboratory and wild," in *International Conference on Image Processing (ICIP)* (IEEE, 2020).
55. D. Yang, V.-T. Peltoketo, and J.-K. Kämäräinen, "CNN-based cross-dataset no-reference image quality assessment," in *International Conference on Computer Vision Workshop (ICCV-W)* (IEEE, 2019), pp. 3913–3921.
56. S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *CVPR* (IEEE, 2020), pp. 3667–3676.
57. D. Li, T. Jiang, and M. Jiang, "Norm-in-norm loss with faster convergence and better performance for image quality assessment," in *International Conference on Multimedia* (ACM, 2020), pp. 789–797.
58. D. Varga, "No-reference image quality assessment with multi-scale orderless pooling of deep features," MDPI J. Imaging **7**, 112 (2021).
59. F. Yu and V. Koltun, "Multi-scale context aggregation by dilated convolutions," in *International Conference on Learning Representations* (2016).
60. C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-ResNet and the impact of residual connections on learning," in *International Conference Learning Representations Workshop (ICLR-W)* (2016).
61. "910. Subjective video quality assessment methods for multimedia applications," P. ITU (International Telecommunications Union Telecommunication Sector, 1999).
62. V. Hosu, B. Goldlucke, and D. Saupe, "Effective aesthetics prediction with multi-level spatially pooled features," in *Conference on Computer Vision and Pattern Recognition (CVPR)* (IEEE/CVF, 2019), pp. 9375–9383.
63. H. Talebi and P. Milanfar, "NIMA: neural image assessment," IEEE Trans. Image Process. **27**, 3998–4011 (2018).
64. A. Paszke, S. Gross, S. Chintala, G. Chanan, E. Yang, Z. DeVito, Z. Lin, A. Desmaison, L. Antiga, and A. Lerer, "Automatic differentiation in pytorch," in *Neural Information Processing Systems (NIPS)* (2017).
65. D. P. Kingma and J. Ba, "Adam: a method for stochastic optimization," arXiv:1412.6980 (2014).
66. L. Celona and R. Schettini, "Blind image quality assessment of authentically distorted images," GitHub (2021) [accessed 8 November 2021], https://github.com/CeLuigi/BIQA4ConsumerPhotographs.