

Blind Multimodal Quality Assessment: A Brief Survey and A Case Study of Low-light Images

Miaohui Wang, *Member, IEEE*, Zhuowei Xu, Mai Xu, *Senior Member, IEEE*, and Weisi Lin, *Fellow, IEEE*,

Abstract—Blind image quality assessment (BIQA) aims at automatically and accurately forecasting objective scores for visual signals, which has been widely used to monitor product and service quality in low-light applications, covering smartphone photography, video surveillance, autonomous driving, etc. Recent developments in this field are dominated by unimodal solutions inconsistent with human subjective rating patterns, where human visual perception is simultaneously reflected by multiple sensory information (e.g., sight and hearing). In this article, we present a unique blind multimodal quality assessment (BMQA) of low-light images from subjective evaluation to objective score. To investigate the multimodal mechanism, we first establish a multimodal low-light image quality (MLIQ) database with authentic low-light distortions, containing image and audio modality pairs. Further, we specially design the key modules of BMQA, considering multimodal quality representation, latent feature alignment and fusion, and hybrid self-supervised and supervised learning. Extensive experiments show that our BMQA yields state-of-the-art accuracy on the proposed MLIQ benchmark database. In particular, we also build an independent single-image modality Dark-4K database, which is used to verify its applicability and generalization performance in mainstream unimodal applications. Qualitative and quantitative results on Dark-4K show that BMQA achieves superior performance to existing BIQA approaches as long as a pre-trained quality semantic description model is provided. The proposed framework and two databases as well as the collected BIQA methods and evaluation metrics are made publicly available.

Index Terms—Low-light image quality assessment, visual-audio quality database, quality semantic description, multimodal learning.

1 INTRODUCTION

STORAGE, transmission, and processing of low-light images are unavoidable [35], especially in smartphone photography, video surveillance, autonomous driving, etc. Imaging in weak-illumination environments can lead to uneven brightness, poor visibility, impaired color, and increased hybrid noise, degrading both user experience and product value [78]. Further, low-light images also pose various challenges to the performance of mainstream vision algorithms, including object detection [13], recognition [92], classification [45], tracking [97], assessment [80], segmentation [14], and enhancement [106]. Therefore, it is essential to develop a reliable objective quality indicator for low-light images, which helps to meet the quality measurement and inspection needs in various industrial products and computer vision tasks.

Low-light blind image quality assessment (BIQA) aims to automatically and accurately estimate objective scores, thereby avoiding multiple obstacles of subjective experiments (e.g., time-consuming, unstable, and non-automated) [78], which plays a significant role in the quality monitoring of industrial products [72]. To address this difficulty, some preliminary models have been developed for the low-light BIQA, such as brightness-guided [88], colorfulness-inspired [79], visibility-induced [78], and comparative learning-based [80]. At the same time, the human visual system (HVS) is

the final receiver of visual signals in the BIQA task [72], [82], [103], and human visual perception is simultaneously reflected by multiple sensory information [54]. However, existing hand-crafted and deep-learned BIQA methods rarely consider multimodal information [55] and are still limited to low-light images alone. As a result, how to utilize multimodal learning to more accurately perform the quality assessment of low-light images is the most fundamental motivation behind this work.

When scoring the quality of visual signals, humans can perceive multiple sensing information (e.g., sight, hearing, etc) at the same time [3], [55], [60]. After acquired exercise, our brains can easily make connections between different modality data and further build a good representation of the characteristics of things [67]. For example, when the image modality is influenced by various low-light noises, other auxiliary modalities are expected to provide supplementary quality description clues, such as ambient sound [54] and semantic visual understanding [77]. Consequently, multimodal BIQA aims to build a visual indicator similar to the manner of human visual perception, and it is expected to learn better quality descriptors to represent human visual perception.

Inspired by the above discussion, we establish an early multimodal BIQA paradigm for low-light images. Considering that there is no low-light BIQA database equipped with multimodal information, we construct the first Multimodal Low-light Image Quality (MLIQ) database. For the image modality, low-light images contain authentic distortions from the steps of image acquisition and processing [78]. For the audio modality, we mainly specify quality-aware principles for generating semantic descriptions of image quality, which are based on the fact that humans are better at

- M. Wang and Z. Xu are with the Guangdong Key Laboratory of Intelligent Information Processing, Shenzhen University, Shenzhen, China.
E-mail: wang.miaohui@gmail.com
- M. Xu is with the School of Electronic and Information Engineering, Beihang University, Beijing. E-mail: maiyu@buaa.edu.cn.
- W. Lin is with the School of Computer Science and Engineering, Nanyang Technological University, 50 Nanyang Avenue, Singapore.
E-mail: wslin@ntu.edu.sg.

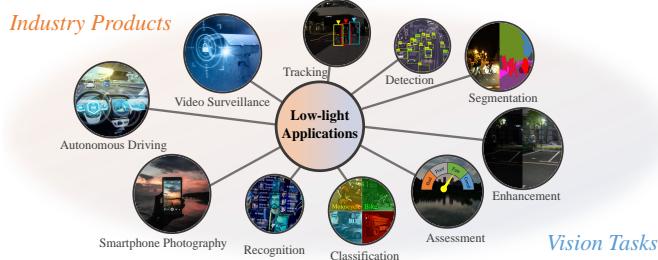


Fig. 1. **Typical applications of low-light image.** Please zoom in the electronic version for better details.

describing quality cognition than giving a quantitative value [96]. Thus, quality semantic description (QSD) contained in audio clips can provide supplementary quality description in the modeling of BIQA.

Further, we develop a unique Blind Multimodal Quality Assessment (BMQA) method to integrate cross-modal features (*i.e.*, image and QSD). The integration of multimodal information helps maintain the representation depth of objective visual signals while broadening the breadth of human visual perception, which can introduce new benefits for the learning of quality descriptors. The expansion of data modalities helps a deep-learned model to enrich low-level embedding features from different perspectives, thereby improving the robustness of the forecasting performance [3]. Extensive experiments validate the effectiveness of the proposed BMQA, which demonstrates the great potential of multimodal learning in blind quality assessment modeling.

The main contributions are five-folds:

- We provide a brief analysis and discussion in various aspects, covering hand-crafted BIQAs (*e.g.*, distortion-specific and general-purpose), and deep-learned BIQAs (*e.g.*, supervised learning-based and unsupervised learning-based). In addition, we also review the exploration of multimodality-based quality assessment in user-perceived quality of experience (QoE).
- Inspired by the human visual perception mechanism, we propose to apply multimodal learning to the BIQA problem by integrating visual and QSD features. To the best of our survey, this is one of the first attempts to explicitly explore low-light quality assessment across different modalities.
- To verify the feasibility of multimodality in the BIQA task, we first construct a new MLIQ database based on low-light images, which contains 3600 pairs of image and audio data. In addition, we have carried out a statistical analysis of the QSD feature, which is helpful to demonstrate human quality cognition.
- Based on the MLIQ database, we further investigate three main modules in multimodal learning, including multimodal quality representation, latent feature alignment, and fusion prediction. To improve the efficiency of deep model training, we develop an effective BMQA method by incorporating both multimodal self-supervision and supervision.
- To demonstrate the applicability of our BMQA, we have also established a new low-light image quality database, namely Dark-4K, which contains only a single image modality. Dark-4K is used to verify

the applicability and generalization performance under the unimodal assessment scenarios. Experiments show that this hybrid learning paradigm guarantees that BMQA achieves state-of-the-art performance on both MLIQ and Dark-4K databases.

2 RELATED WORK

According to the extraction manner of quality features, we mainly review some representative BIQA methods from two aspects: 1) hand-crafted and 2) deep-learned BIQAs. In addition, we also introduce recent visual-audio quality assessment methods. For a detailed overview of BIQA, the reader is referred to [100].

2.1 Hand-crafted BIQA Methods

Hand-crafted BIQAs are usually based on the feature extraction of expert and engineering experience. Since hand-crafted BIQAs require less deployment environment (*e.g.*, database size, hardware platform, computing power, *etc.*), they are highly achievable and easy to be deployed. Based on the concrete application, existing BIQAs can be divided into distortion-specified and general-purposed ones.

2.1.1 Distortion-specific BIQAs

Distortion-specific BIQAs measure an image quality by considering both degradation manners and distortion types of a special application as shown in Fig. 1. In this section, we briefly review some representative applications.

Screen Content. The development of computer-generated technology has promoted the widespread of screen content (SC) visual signals [53], and further drives the requirement of SC-based BIQA. Based on the distinct features (*e.g.*, sharp edge, thin line, limited color, and high contrast) of SC images, many representative BIQAs have been proposed, such as structural feature-induced [24], brightness and texture-driven [19], and region division-based [2], [107].

High Dynamic Range (HDR). HDR-based BIQAs measure the quality degradation mainly caused by tone mapping or multiple exposure fusion. Some BIQAs have considered representative quality descriptors, such as gradient [34], color [50], and brightness [29]. With the continuous development of HDR, the use of new technologies (tone-mapping or exposure fusion algorithm) has improved HDR imaging results. It remains to be verified whether existing BIQAs can accurately measure the distortion introduced by new HDR technologies.

Omnidirectional Stereo. The popularity of immersive multimedia applications has promoted the development of stereoscopic-omnidirectional BIQA tasks. One of the main quality feature extraction paradigms is based on multi-view decomposition [27], [28], [110]. However, in addition to the quality of the visual signal itself, there are many factors that affect the perceived quality of virtual reality, such as viewing behavior [71] and camera motion [91]. Therefore, there are still many quality descriptors to be explored for the stereoscopic-omnidirectional scenario.

Other Applications. Besides the application mentioned above, there are other promising distortion-specific BIQAs designed for emerging computer vision tasks, such as super-resolution [109], segmentation [66], dehazing [6], deraining [87], aesthetic [15], encryption [84], light-field [73], underwater [108], 3D geometry [81], *etc.* Due to the length of

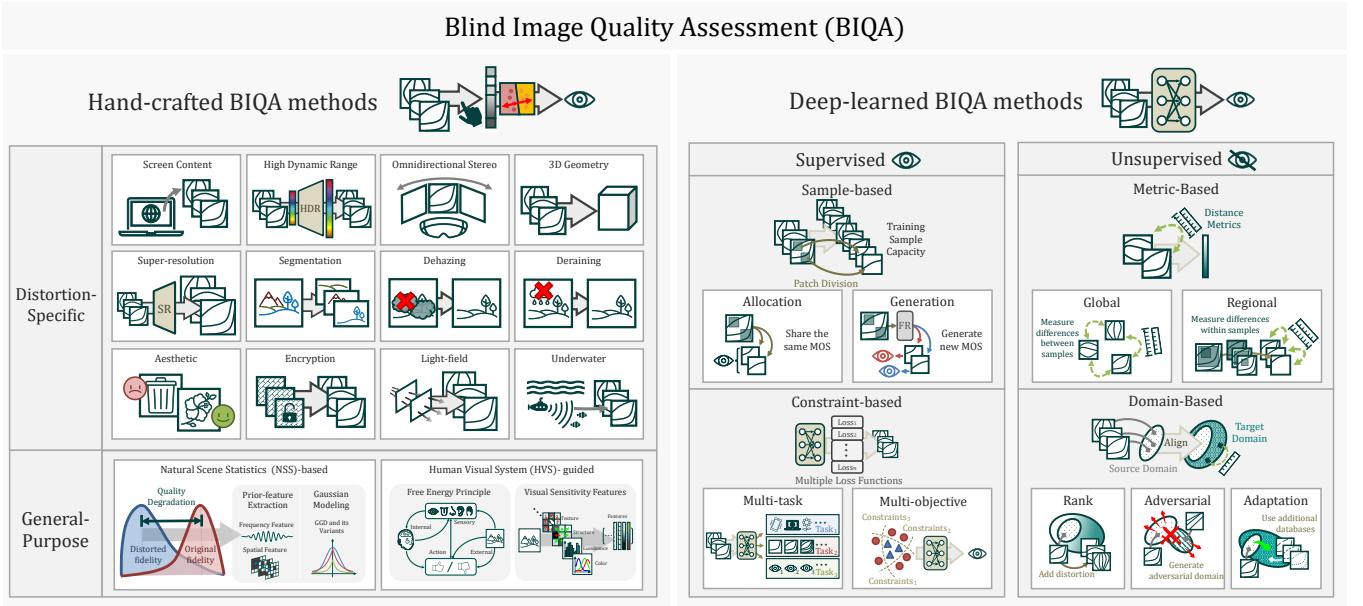


Fig. 2. **A concise illustration of blind image quality assessment (BIQA).** Representative methodologies are briefly demonstrated from the perspective of hand-crafted BIQA methods (*e.g.*, distortion-specific and general-purpose) and deep-learned BIQA methods (*e.g.*, supervised learning-based and unsupervised learning-based).

the article, we will not discuss other types of distortion-specific BIQAs here. A brief illustration of the related work is organized as shown in Fig. 2.

2.1.2 General-purpose BIQAs

Unlike distortion-specific BIQAs, general-purpose methods are not designed for the typical image applications described above. These BIQAs usually utilize some common quality-aware features to quantify image distortion, which can be roughly divided into natural scene statistics (NSS)-based and HVS-guided BIQA methods.

NSS-based. NSS-based BIQAs are based on the assumption that high-fidelity images obey some kind of prior statistical characteristics that will be altered by quality degradation [58]. For instance, early BIQAs quantify these changes in the frequency domain, such as wavelet transform [58] or discrete cosine transform [64]. Considering the computational complexity of domain transformation, some BIQAs directly select descriptive quality-aware features in the spatial domain, such as normalized luminance [56], gradient magnitude [93], and local binary pattern [44]. To better model the NSS characteristics, parametric models have also been developed, such as multivariate generalized Gaussian density (GGD) [57] and asymmetric GGD [102]. Existing NSS-based BIQAs are rarely able to fully consider the distortion characteristics of different image distortion types, so it is difficult for such methods to achieve optimal performance in practical applications.

HVS-guided. Due to the fact that the ultimate recipients of visual signals are humans, it is significant to exploit the perception characteristics of the HVS in the design of image quality indicators. There are some representative HVS-guided BIQAs, including free-energy principle driven and visual sensitivity-based BIQAs. The free-energy principle interprets the quality perception of input visual signals as an active inference process [101]. On the other hand, the visual characteristic-based BIQAs mainly leverage some

visual sensitivity features, such as luminance [38], texture [42], and color [20]. However, one of the main limitations of HVS-guided BIQAs is that general-purpose distortions can be very complex and affected by various factors, and it is difficult to be adequately characterized by simple HVS models and visual features.

Existing general-purpose BIQAs can achieve satisfactory performance on synthetic databases, but the performance is relatively limited when faced with images captured in real-world weak-illumination scenarios. Meanwhile, with the refinement of industrial products or services, the imaging quality of vision sensors in specific scenes (*e.g.*, taking pictures at night) is more likely to be used as a product value indicator [78]. Thus, a low-light image database with authentic distortions has important practical significance in both industry and academia.

2.2 Deep-learned BIQA Methods

Deep-learned BIQAs [80] directly learn quality features from distorted images in an end-to-end manner. In contrast to the hand-crafted BIQAs, these methods automatically optimize quality forecasting models which have shown promising performance. Deep-learned BIQAs can be usually divided into supervised learning-based and unsupervised learning-based BIQA methods. It is worth noting that there are other learning types (*e.g.*, reinforcement learning [23], [65]), which are somewhat less commonly used in the deep-learned BIQA task.

2.2.1 Supervised Learning-based BIQAs

The learning objective of a supervised learning-based BIQA is to minimize the difference between the predicted score and the subjective mean opinion score (MOS) value. The current supervised learning BIQAs focus on solving the problem of insufficient training samples, and these methods can be mainly categorized into sample-based and constraint-based ones.

Sampled-based. Sample-based BIQAs are mainly based on expanding the capacity of training samples, which usually utilize patch-level quality features to predict an image-level score. Existing sample-based BIQAs mainly consist of 1) annotation allocation-based and 2) annotation generation-based methods.

Allocation-based BIQAs directly share the same image-level annotations to all patches in a given image [30]. Subsequent improvements have established the correlation between patch-level features and image-level scores, such as weighted features [89], weighted decisions [4], and voting decisions [47]. Recent attempts have also extended an input image into multi-scale patches and learned a general feature representation [31]. However, uncertainty in image content makes it difficult for these BIQAs to capture highly non-stationary feature representations due to local content variations and complicated correlations between content and distortion [47].

Generation-based BIQAs usually learn a supervised learning model via patch-level scores. Due to the lack of a region-level based MOS database, early BIQAs employed full-reference models to generate quality scores for image patches [32]. However, the forecasting performance is highly dependent on the adaptability and correlation between the full-reference model and the target BIQA task [104]. Recent progress has benefited from the substantial human involvement in building new databases with image and patch-level scoring labels [99]. At the same time, this shows that there is a huge cost challenge to establish a large enough database for practical quality inspection applications.

Constraint-based. Constraint-based BIQAs optimize multiple loss functions simultaneously in supervised learning, which can be roughly divided into 1) multi-task and 2) multi-objective ones. Multi-task based BIQAs often train additional goals which are highly associated with image quality, such as distortion types [47], error maps [33], natural scene categories [94], and content attributes [104].

Multi-objective based BIQAs simultaneously optimize multiple constraints (or regularization terms) to improve the model training [70], such as employing multiple loss functions for multi-scale supervision [86], introducing new normalization embeddings into the objective function [36], and using additional constraints to adjust initialization parameters [111]. Recent developments also incorporate constraints learned from other databases, such as using incremental learning to measure correlation across databases [48], using uncertainty prediction to rank the fidelity across databases [105], and using continuous learning to measure similarity across databases [103].

2.2.2 Unsupervised Learning-based BIQAs

Unsupervised learning-BIQAs extract latent embedding features without using any ground-truth MOS labels, which are further employed to measure the quality difference between each training sample based on its latent feature. While massive hand-crafted labels are not required to extract the latent embedding features in unsupervised learning, the less explicit objective function makes the design of training a model more challenging. Existing studies can be categorized into metric-based and domain-based BIQAs.

Metric-based. Metric-based BIQAs employ some widely used distance measurements (*e.g.*, cosine similarity, Wasserstein distance, *etc.*) to extract latent embedding features which are used to measure the difference (*e.g.*, distortion type, distortion level, and content category) between the current image sample and the other samples in the training database. Recently, various distortion descriptors have been developed to extract global embedding features, such as distortion type-based discriminative learning [49], distortion level-based contrastive learning [83], and content category-based similarity learning [59]. In general, these metric-based BIQAs treat training samples as mutual quality references and maximize the difference in quality features.

Domain-based. Domain-based BIQAs commonly design domain alignment constraints, and measure the quality difference between each sample in a source domain based on the error metric defined in a target domain. Early domain-based methods directly add different levels of synthetic distortion and learn to rank them [43], [46]. Subsequent domain adversarial BIQAs employ a generator to build the target domain and then guide the source domain for adversarial learning [39], [63]. Recently, domain adaptation BIQAs use additional large databases to construct the target domain and guide the source domain to learn the rules of quality description in the target domain [9], [12]. Domain-based BIQAs usually require strict assumptions, which make them difficult to meet the model requirements when the distortion type of a testing image is unknown.

2.3 Multimodality Assessment and Motivation

2.3.1 Visual-audio Quality Assessment

Visual-audio quality assessment refers to a quantitative evaluation of the user-perceived QoE, which jointly considers cross-media quality. It can be divided into degradation-based and perception-based methods.

Degradation-based. Existing degradation-based methods commonly measure the user-perceived quality score of visual and audio independently, and then design a combination rule to predict the quality score. This paradigm is based on that the degradation of video and audio only depends on signal capture equipment [61], compression, transmission and terminals, but they do not degrade each other (*i.e.*, video degradation does not cause audio degradation). The combination rule of video and audio mainly considers addition [25], multiplication [85], vote [76], weighted Minkowski [51], *etc.* Due to the fact that degradation-based methods ignore the interaction between cross-media information, it is difficult to simply combine individual quality scores to accurately estimate the user-perceived QoE.

Perception-based. The perception-based methods mainly learn a joint perceptual feature space, where video and audio features can both be represented and combined to predict the quality score. This paradigm is inspired by psychophysical experiments (*e.g.*, the McGurk effect [52]), where cross-media perceptual interaction has a greater impact on the user-perceived quality. For example, user-perceived video quality is more favorable when paired with high-fidelity sound than when used alone, but user-perceived audio quality is more favorable when used alone than when paired with high-quality videos [69]. Therefore,

the user-perceived QoE should be considered based on the combination of perceptual features to capture the above cross-media biases. Existing explorations mainly focus on the feature combination of spatial perception [7], temporal perception [98], and spatio-temporal perception [54].

2.3.2 Motivation

As discussed in [54], [60], there are many quality evaluation databases separately established for video and audio, respectively. However, very few databases have been established containing both the video and audio subjective labels. Among these visual-audio databases, the closest to our MLIQ is the LIVE-SJTU audio and video quality assessment database (A/V-QA) [54]. A/V-QA consists of 14 video sequences selected from the *Consumer Digital Video Library* (CDVL) [60] with the original stereophonic soundtrack. For each video sequence, there are a total of 24 quality degradation versions: the video part is compressed or sampled by 4 quantization levels, while the audio part is compressed at 3 bit-rate levels.

Unlike the A/V-QA, MLIQ consists of low-light image samples and their QSD-based audios:

- 1) The audio content of MLIQ reflects the image quality in terms of semantic visual understanding [75], while that of A/V-QA can be silence, noise, pure music, pure speech, or speech over background sound.
- 2) MLIQ contains various types of authentic distortions, while A/V-QA only contains synthetic compression and sampling distortions.
- 3) MLIQ is established to study the impact of auxiliary modalities on the BMQA task, while A/V-QA is constructed to study the impact of visual and audio distortion on the overall QoE. In other words, the MOS labels of MLIQ depend only on images, while the MOS labels of A/V-QA depend on both images and audio clips.

The differences between existing visual-audio quality indicators and our BMQA are obvious:

- 1) The former aims to predict the joint user-perceived quality of cross-media, while the latter aims to predict the visual quality of a given image.
- 2) In current visual-audio quality assessment, other modalities (*e.g.*, audio, text, *etc*) are seldom exploited to assist in predicting more accurate visual quality scores. Our BMQA focuses on studying the impact of auxiliary modalities on image quality forecasting.
- 3) Existing visual-audio quality indicators rarely consider deep multimodal learning, such as multimodal feature extraction, alignment, fusion, *etc*. Our BMQA investigates various cross-media learning modules, from deep quality representation to quality fusion prediction.

In addition, it is worth noting that humans are better at measuring image quality by semantic description rather than quantitative score [96], and hence QSD can be a very useful modality in the modeling of BMQA. The reasons why this investigation uses the audio modality as the QSD information carrier are based on: 1) The speech-to-text feature conversion is very convenient and mature; 2) The audio

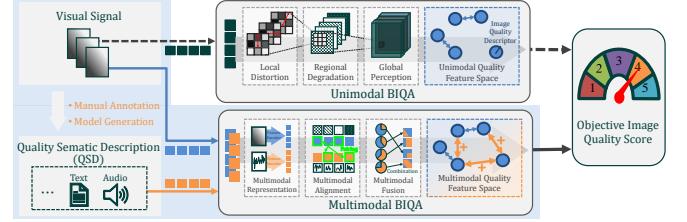


Fig. 3. Relationship between unimodal and multimodal BIQAs. Humans are better at perceiving image quality through semantic descriptions rather than quantitative values, which reveals that quality semantic description (QSD) is a very useful modality for BIQA modeling.

modality itself can carry a wider range of quality features (pitch, timbre, rhythm, *etc*), which retains the possibility to study other quality features in the future.

Furthermore, some may raise the challenge that there are few audio information scenarios when performing BIQA tasks. We argue that our investigation focuses on how additional modality information (*e.g.*, semantic visual understanding in terms of QSD), if present, affects the modeling of BMQA. More importantly, QSD can be encapsulated in audio, text, or other forms. Semantic visual understanding methods, such as image captioning [90], can be directly used to generate text-based QSD modality information. Based on our ML IQ database, we can pre-train a QSD-induced captioning model, which allows our BMQA scheme to be used seamlessly in any existing unimodal BIQA task. Therefore, this lack of audio scenario is not a problem for our investigation. The relationship between unimodal BIQAs and our BMQA is illustrated in Fig. 3.

Finally, multimodality-driven BIQAs are still in their early stage, but BMQA provides a new and promising perspective. On one hand, the homogeneity of multimodal data shows that training information can be supplementary or shared, which facilitates the learning of highly descriptive quality features. On the other, the heterogeneity of multimodal data can expand the width and depth of training information, which is expected to improve the forecasting performance. To this end, it is highly desired to develop a special multimodal quality indicator for low-light images.

3 MULTIMODAL LOW-LIGHT IMAGE QUALITY (MLIQ) DATABASE CONSTRUCTION

In this section, we describe how to construct our ML IQ database. The established database contains RGB images with the subjective quality scores and QSD-based audio clips with the quality semantic description (QSD) as shown in Fig. 4.

3.1 Multimodality Construction

3.1.1 MOS-based Image Modality

Natural Night-time Image Database (NNID) [88] is the latest publicly available no-reference low-light image database. It contains 448 different visual scenes with a total of 2240 samples, covering daily life recording, intelligent transportation, surveillance, city light show, aerial photography, and many other application scenarios. The characteristics of the NNID database are very suitable for our experimental requirement. To facilitate the acquisition

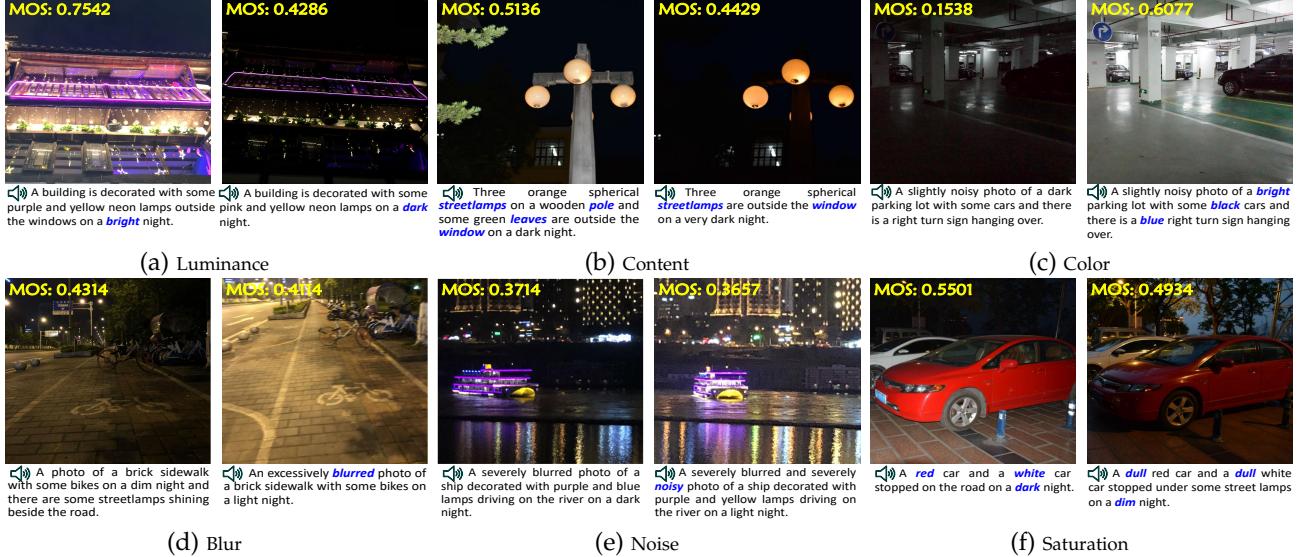


Fig. 4. **Examples of some Image-QSD-MOS pairs on our MLIQ database.** We provide some representative examples for analysis and discussion of the QSD features, including luminance, content, color, blurry, noise, and saturation. Please zoom in the electronic version for better details.

of low-light image data for MLIQ, we adopt it as part of our benchmark image source.

Furthermore, we adopt another 1360 low-light images captured by two new devices, *Canon EOS 6D* and *Huawei Mate 30 Pro*, to expand the coverage of weak-illumination scenes. MLIQ consists of a total of 3600 low-light image samples. These low-light samples are captured in real-world environments (*e.g.*, indoors and outdoors) with a total of five different mobile devices. One device captures a visual scene with five different settings. These five settings are allowed to be different for different scenarios. The resolution of each low-light sample ranges from 512×512 to 2048×2048 . Therefore, MLIQ is the largest no-reference low-light database, covering various scenes, large volumes, complex noise, diverse devices, and authentic distortion.

To obtain the MOS label for each low-light image on MLIQ, we have conducted a subjective experiment. Following ITU-R BT.500-14 [5] and NNID [88], we take a single stimulus and build a graphical user interface to perform the experiments. *Skyworth 28U1* is used, and the viewing distance is approximately three times of the image height. We have invited 26 participants including 16 females and 10 males (between the ages of 18 and 35). The participants are asked to score the image quality based on 11 discrete quality scores ranging from 0 to 1 with a step length of 0.1. The rated score will finally correspond to five quality levels, including [0, 0.1] for *bad*, [0.2, 0.3] for *poor*, [0.4, 0.5] for *fair*, [0.6, 0.7] for *good*, and [0.8, 1.0] for *excellent*.

A statistical analysis of our MLIQ is illustrated in Fig. 5. Figs. 5 (a)-(c) provide the overall statistical data of shooting device, image resolution, and content application scenario, respectively. Fig. 5 (d) reports the histogram distribution of MOS values. As seen, MOS values span the entire quantified range of visual quality with sufficient and fairly uniform samples at each level. This shows that our MLIQ database covers the entire range of visual quality (from poor to excellent), and also exhibits a good separation of the perceptual quality. In addition, Fig. 5 (e) reports the 95% confidence intervals obtained from the mean and standard deviation of the rating score for each image as the consistency eval-

uation, where the confidence interval is mainly distributed between 0.11 and 0.17. It indicates that all observers have reached a high agreement on the perceptual quality of low-light images. Therefore, the proposed MLIQ database is used as a ground-truth for the performance evaluation of objective quality indicators.

3.1.2 QSD-based Audio Modality

A simple and effective way to represent human visual perception is to ask subjects to describe and record their semantic visual understanding [75], [77]. The QSD information can provide additional quality assessment clues: On one hand, free verbal descriptions of visual understanding provide potentially the richest QSD source, since language is the most flexible means of human communication. On the other, verbal descriptions help to avoid personal information bias, as verbal descriptions are far less likely to be inconsistent than consistent [26], [75].

By synthesizing some previous work, we design two QSD principles based on the perception mechanism of the HVS. It can exhibit feed-forward visual information extraction and aggregation from the retina (*i.e.*, *intuitive visual perception*) to the primary visual cortex (*i.e.*, *empirical visual perception*) [95]. These principles are used to guide annotators in generating their verbal descriptions.

Intuitive Visual Perception. This principle is inspired by previous physiological and psychological experiments on the HVS, including saliency detection and just noticeable difference. It is closely related to early vision and focuses on the relationship between optical stimuli from visual observation and the HVS experience [75]. Intuitive vision mainly focuses on some basic data characteristics, covering overall brightness, color status, texture, salient objects, *etc.* For instance, the verbal description in Fig. 4 (c) contains the QSD features, such as the luminance information '*bright*', the color information '*black*' and '*blue*', and the observed object information '*car*' and '*sign*'.

Empirical Visual Perception. This principle is inspired by modern theoretical studies in visual theory that embraces

empiricism as the dominant paradigm of human perception construction, such as Gregory's theory [22]. These studies demonstrate that while part of what we perceive comes from objects in front of us, another part (and possibly a larger part) always comes from our brains. The empiricism principle is closely related to late vision and focuses on exploring how the viewpoints of observers are involved in quality cognition. Empirical vision mainly involves some common real-world knowledge as well as empirical visual understanding, and highlights the possible real-life scenarios of low-light observations [21]. For instance, subjects use '*driving*' rather than '*sitting*' for '*ship*' as shown in Fig. 4 (e).

There are two main challenges in obtaining the QSD labels: On one hand, there is still a lack of research on how to construct the QSD features between the audio and image data. On the other, various audio components convey the QSD information, such as timbre, pitch, rhythm, fluency, etc. It is necessary to avoid information deviations coming from different audio components as much as possible. Thus, we construct the QSD-based audio clips in the manner of scene dictation which is based on two reasons: 1) Scene dictation presents human observations and is dominated by semantic information, which is at the cognitive level and thus contains little subjective emotion. 2) Scene dictation also helps to avoid information bias due to the speaker's voice or personal pronunciation preference in later analysis. In the experiments, speakers are asked to provide a meaningful and concise verbal description for each low-light image. For the content of each dictation, the following requirements need to be met:

- For images with salient objects (*e.g.*, the ship in Fig. 4 (e)), trying to describe all important objects in the image content. For images with salient scenes (*e.g.*, the parking scene in Fig. 4 (c)), trying to describe the overall environment. For images without any salient content (*e.g.*, the building with many small objects in Fig. 4 (a)), trying to describe attractive content part, including objects and scenes.
- Trying to describe the overall brightness by using the relevant QSD features, such as '*bright*', '*light*', '*dim*', or '*dark*'.
- Trying to describe the main attributes of each object, such as color, brightness, texture (*e.g.*, '*wooden*' pole in Fig. 4 (b) and '*brick*' sidewalk in Fig. 4 (d)), etc.
- Trying to describe the quality cognition by using the relevant QSD features, such as '*colorful*', '*vivid*', '*blurred*', '*noisy*', etc.

In the experiments, 14 male and 10 female trained participants are invited to provide a meaningful and concise verbal description for each low-light image. Next, we record the QSD clip in an environment below 25 dB. After the recording procedure, we invite another 21 judges to rate the audio clips, where each judge is asked to vote for the most representative speech that is most clearly descriptive and does not carry emotion. The highest-voted recorded dictation is selected as the QSD modality. Finally, we collect 3600 audio clips, where the length of the QSD labels ranges from 6 to 45 words.

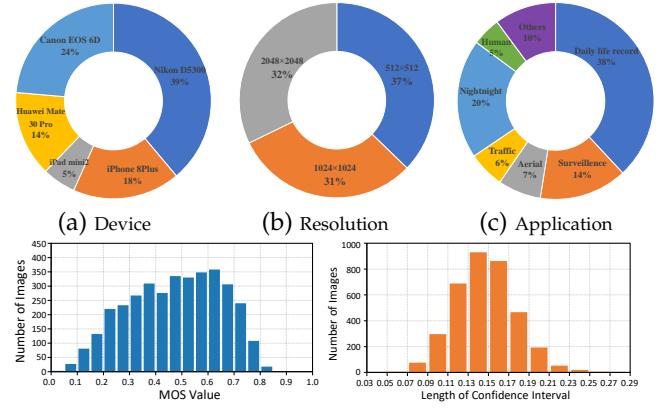


Fig. 5. A statistical analysis of the proposed multimodal MLIQ database: (a) shooting device, (b) image resolution, (c) application scenario, (d) MOS distribution, and (e) confidence interval.

3.2 Database Analysis

In this section, we analyze correlations between images, QSD features, and quality scores (*i.e.*, MOS) on the MLIQ database. Specifically, the image modality represents a visual stimulus, and the corresponding QSD information represents a subjective quality cognition and understanding of the associated image modality. Based on the statistical analysis of MLIQ, we attempt to capture the underlying connection between visual signals and verbal descriptions. We conduct the statistical analysis based on brightness, content and color, and then discuss other factors affecting quality perception on low-light images.

3.2.1 Luminance

Low-light images mainly suffer from insufficient or uneven brightness. The quality level of low-light images is sensitively dependent on visual brightness perception. The QSD information contains quality clues (*i.e.*, keywords) that describe the luminance status, which can effectively provide supplementary information. Therefore, it is significantly meaningful to explore the relationship between brightness and quality.

We start by figuring out how QSD represents the image quality from the view of brightness. The luminance-based QSD features, covering '*dark*', '*dim*', '*light*', and '*bright*', represent the illumination condition. We calculate the histogram of the luminance QSD feature on the entire database, as shown in Fig. 6 (a). As seen, the MOS value corresponding to '*dark*', '*dim*', '*light*', and '*bright*' is concentrated around 0.2 to 0.4, 0.4 to 0.5, 0.5 to 0.6, and 0.6 to 0.8, respectively. Intuitively, the histogram of each luminance QSD feature should obey an independent Gaussian distribution. Therefore, we adopt a Gaussian function to fit the histogram of '*dark*', '*dim*', '*light*', and '*bright*', and the Gaussian centers are 0.2797, 0.4338, 0.5744, and 0.6745, respectively. The above observations suggest that the luminance QSD features have a strong relationship with the image quality.

Next, we analyze the relationship between images, QSD features, and quality scores based on brightness. We report the stacked histogram of the length of verbal at various luminance conditions in Fig. 6 (b). As seen, '*dim*' and '*dark*' represent low luminance, which tend to have shorter verbal lengths. This is consistent with our experience that people

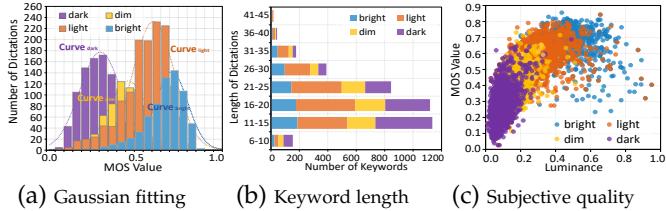


Fig. 6. **Statistics of the luminance QSD feature:** (a) histogram of the luminance keywords and Gaussian fitting, (b) histogram of the length and number of luminance keywords, and (c) scatter distribution of the luminance keywords and MOS values.

often have difficulty describing very dark scenes with long verbal descriptions.

In addition, we calculate the average luminance value for each image as the objective luminance level and report the scatter plot of the corresponding quality score as shown in Fig. 6 (c). We further mark colors for each plot based on the luminance keyword contained in the corresponding QSD. It can be observed that as the luminance level increases, the quality score generally increases as shown in different scatter colors. It suggests that the luminance QSD feature is an efficient representation of visual quality perception.

3.2.2 Content

Due to insufficient exposure, low-light distortions usually result in incomplete or unclear visual quality, further leading to an annoying visual experience. The QSD modality contains verbal descriptions of observed objects, which can effectively provide auxiliary information on which objects the visual attention is focused on. Therefore, we explore the relationship between image quality and content.

The reduction of observed objects is often reflected in the reduction of object descriptions in the QSD modality, as shown in Fig. 4 (b). Based on this observation, we count the quantity of observed objects and report the corresponding MOS values. Fig. 7 (a) consists of stacked column charts and scatter plots, covering the number of observed objects ranging from 1 to 5. For each number of observed objects, the stacked column chart reports the image number at each luminance level, including 'dark', 'dim', 'light', and 'bright', respectively.

Based on the statistical data of image content, we can draw some interesting conclusions: 1) The curve in Fig. 7 (a) shows that the quality score tends to be higher as the quantity of observed objects increases. This may indicate that images with better visual quality usually contain more identifiable observed objects. 2) The stacked histogram in Fig. 7 (a) shows that when the luminance levels get lower, the number of observed objects decreases. 3) The quality score increase caused by the quantity increase of observed objects is small (0.0657 from 1 to 5), which indicates that it is difficult to sensitively reflect visual experience via the quantity of observed objects. One possible reason may be that low-light distortion tends to lose detail rather than salient objects, while the quality score depends more on the salient content itself.

3.2.3 Color

Low-light distortion tends to exhibit low color contrast and low saturation. The observed colors may effectively

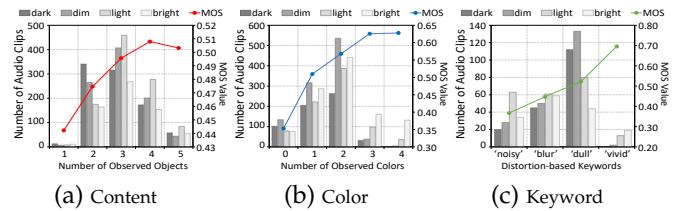


Fig. 7. **Statistics of other QSD features:** (a) content, (b) color, and (c) keyword. The number of related images is measured by the main ordinate (placed on the left-hand side), while the average MOS result (dots connected by a solid line) is measured by the secondary ordinate (placed on the right-hand side).

provide useful QSD information on visual perception responses. Therefore, we investigate the relationship between image quality and observed colors.

The impairment of observed colors is reflected as the reduction of color descriptions, as shown in Fig. 4 (c). Inspired by this, we count the number of color QSD features and report the corresponding MOS values. Fig. 7 (b) consists of stacked column charts and scatter plots, covering the number of color words ranging from 0 to 4. For each number of color QSD features, the stacked column chart reports the image quantity at each luminance level, including 'dark', 'dim', 'light' and 'bright', respectively.

Based on the statistical data of image color, we can also draw some interesting conclusions: 1) The curve in Fig. 7 (b) shows that the quality score tends to be higher as the number of color QSD features increases. This may indicate that images with a better visual perception experience usually contain more recognizable colors, as shown in Fig. 4 (c). 2) The stacked histogram in Fig. 7 (b) shows that when the visual perceptual luminance gets lower, the number of observed colors tends to be lower. 3) The quality score increases caused by the number increase of observed color is large (0.2747 from 0 to 4). This may indicate that the quantity of observed color can sensitively represent the quality experience of low-light distortion.

3.2.4 Other Factors

Low-light photography is also often affected by many other factors, including blurring, heavy noise and low saturation [11]. A low-light image may get blurred by the camera shake if it is set to a long exposure time, as shown in Fig. 4 (d). The increase in light sensibility reduces the signal-to-noise ratio while increasing the exposure, as shown in Fig. 4 (e). In addition, both underexposure and overexposure significantly affect color saturation, which further affect the visual experience.

Considering that the QSD label may contain some keywords that directly describe these degradation features, we collect the distortion-based keywords and report the corresponding MOS values. Fig. 7 (c) shows stacked column charts and scatter plots, covering the distortion-related keywords of 'blur', 'noisy', 'dull', and 'vivid'.

Based on the above statistical data, we can draw some interesting conclusions: 1) QSD features such as 'blur', 'noisy', and 'dull' represent poor visual experience, while 'vivid' represents good visual experience. 2) The proportion of 'noisy' is large in the 'bright' luminance level, which indicates that noises in low-light images are more easily perceptible. 3) The proportion of 'blur' is similar under different luminance levels, which indicates that blur is not closely related to

lumination. 4) The proportion of ‘dull’ is large in the ‘dark’ and ‘dim’ luminance levels, while the proportion of ‘vivid’ is large in the ‘light’ and ‘bright’ luminance levels. These observations are consistent with the fact that human eyes prefer highly saturated colors, as shown in Fig. 4 (f).

4 PROPOSED BLIND MULTIMODAL QUALITY ASSESSMENT (BMQA) FRAMEWORK

Based on the proposed MLIQ database, we design a unique deep-learned BMQA as shown in Fig. 8. We address the main challenges of multimodality learning in the BIQA task, including feature representation, alignment, and fusion. Finally, we will describe our learning mechanism.

4.1 Multimodal Quality Representation

In this section, multimodal quality representation refers to extracting and integrating effective features that take the advantage of the supplementary quality description clues. Due to the fact that heterologous data has significantly different characteristics, this indicates a large difference between image and QSD spaces [67]. Therefore, we design two different feature quality representations for them.

4.1.1 QSD-based Audio Quality Representation

In this subsection, we explore the quality cognition in terms of the QSD feature, which mainly consists of three stages.

First, we need to avoid background noises and transient disturbances in an audio clip, which would otherwise spoil the human experience of perceptual quality [17]. Fortunately, the environment preparation of the audio recording in Sec. 3.1.2 allows us to ignore these effects. Therefore, the QSD-based audio quality representation \mathbf{F}_{aud} can be defined by

$$\mathbf{F}_{aud} = \mathcal{F}_{aud}(X_{aud}; \theta_{aud}), \quad (1)$$

where \mathcal{F}_{aud} is composed of a speech recognition module and a QSD feature extractor, X_{aud} denotes an input audio clip, and θ_{aud} represents the corresponding weights.

Second, we extract the QSD information, which is a symbolic representation of linguistic words. Since the speech recognition technology has achieved satisfactory results, we will not expand and describe it in detail. In the experiments, we employ an existing speech recognition module [1], $\mathcal{F}_{aud2qsd}$, to transcribe the audio clip into spoken words. $\mathcal{F}_{aud2qsd}$ can be formulated as:

$$< x_1, \dots, x_N > = \mathcal{F}_{aud2qsd}(X_{aud}; \theta_{aud2qsd}), \quad (2)$$

where $\theta_{aud2qsd}$ denotes the parameter weight of $\mathcal{F}_{aud2qsd}$, and $< x_1, \dots, x_N >$ denotes the output spoken words with a length of N .

Third, we further extract the latent QSD feature from spoken words. Given a word $x_k \in < x_1, \dots, x_N >$, we adopt \mathcal{F}_{qsd} to extract the semantic quality feature for each word. In BMQA, \mathcal{F}_{qsd} can be formulated as follows:

$$\begin{aligned} \mathcal{F}_{qsd}(x_k) = & \sigma[\mathcal{F}_{lp}^1(x_n, < x_1, \dots, x_n, 0, \dots >; \theta_{lp}^1)] \\ & \oplus \mathcal{F}_{lp}^2(x_n, < x_1, \dots, x_n, 0, \dots >; \theta_{lp}^2), \end{aligned} \quad (3)$$

where \mathcal{F}_{lp}^1 and \mathcal{F}_{lp}^2 denote two linear projections (e.g., fully connection or attention projection), θ_{lp}^1 and θ_{lp}^2 denote the corresponding weights, σ denotes a nonlinear activation function (e.g., rectified linear unit or gaussian error linear units), and \oplus denotes a fusion operation (e.g., addition or concatenation).

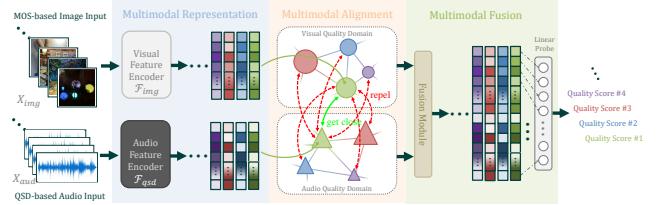


Fig. 8. Overall learning framework of the proposed BMQA. It consists of three key modules, including multimodal feature representation, latent feature alignment, and fusion prediction.

corresponding weights, σ denotes a nonlinear activation function (e.g., rectified linear unit or Gaussian error linear unit), and \oplus denotes a fusion operation (e.g., addition or concatenation). Currently, there is still no consensus on the efficient semantic processing [68]. The main reason is that different methods build different word associations (*i.e.*, different settings of \mathcal{F}_{lp}^1 , \mathcal{F}_{lp}^2 , σ and \oplus), and they show different advantages in natural language processing (NLP).

Finally, the association of all words in a dictation will contribute to the feature representation of \mathbf{F}_{aud} . In BMQA, it is generated by a linear combination of all word features:

$$\mathbf{F}_{aud} = \mathcal{F}_{lc}[\mathcal{F}_{qsd}(x_1), \dots, \mathcal{F}_{qsd}(x_N); \theta_{lc}], \quad (4)$$

where \mathcal{F}_{lc} represents a fully-connected layer and θ_{lc} denotes the weights of \mathcal{F}_{lc} . In the experiments, we select 3 representative networks as the backbone of \mathcal{F}_{qsd} , including Bag-of-Word (denoted as BoW), recurrent neural network (denoted as RNN), and Transformer (denoted as TransF).

4.1.2 MOS-based Image Quality Representation

Recent work has demonstrated the feasibility and superiority of deep neural network (DNN) in many vision tasks. Inspired by this, we obtain the image quality representation \mathbf{F}_{img} by a deep-learned mapping:

$$\mathbf{F}_{img} = \mathcal{F}_{img}(X_{img}; \theta_{img}), \quad (5)$$

where \mathcal{F}_{img} denotes an image feature encoder, X_{img} denotes an input image, and θ_{img} represents the corresponding weights.

Existing deep-learned image quality representation [18] mainly relies on the extraction and integration of features in the spatial domain. \mathcal{F}_{img} is usually designed as a stack of multiple layers of DNN blocks. Specifically, let \mathcal{F}_{img}^l denote a basic DNN block in the l -th layer of \mathcal{F}_{img} . \mathcal{F}_{img}^l usually consists of sequential linear computations and nonlinear activation functions in BMQA, which can be formulated as:

$$\mathcal{F}_{img}^l(X_{img}^l) = \sigma_l[\mathcal{F}_{lp}^1(X_{img}^l; \theta_{lp}^1)] \oplus \mathcal{F}_{lp}^2(X_{img}^l; \theta_{lp}^2), \quad (6)$$

where \mathcal{F}_{lp}^1 and \mathcal{F}_{lp}^2 denote two linear projections (e.g., fully connection, attention projection, or convolution), θ_{lp}^1 and θ_{lp}^2 denote the corresponding weights, σ denotes a nonlinear activation function (e.g., rectified linear unit or gaussian error linear units), and \oplus denotes a fusion operation (e.g., addition or concatenation).

Recent studies demonstrate that different backbones of deep models result in different learning capabilities. It indicates that network architectures may have important impacts on the extraction of \mathbf{F}_{img} . Therefore, we select 5 representative networks as the backbone of \mathcal{F}_{img} , including VGG, ResNet (denoted as RN), EfficientNet (denoted as EN), Vision-in-Transformer (denoted as ViT), and ConvNeXT (denoted as CNXT).

4.2 Multimodal Quality Alignment

In BMQA, multimodal quality alignment refers to finding the corresponding quality representation relationship between two modalities. The image and QSD are obtained for the same scene, and the quality description can be highly consistent [8]. For example, the QSD keywords can directly indicate the image regions that the visual attention focuses on [40], thus improving the learning performance of an image feature encoder.

Existing methods build the multimodal alignment by designing constraints across different modalities (called cross-modal constraints [3]). In the construction of training objective, it is necessary to define a special metric to measure the difference between two modalities. If two modalities come from different perspectives of a single sensor, the metric is defined as an absolute value error, such as mean absolute error (MAE) and mean square error (MSE). If two modalities come from different sensors, the metric is defined as a relative value error, such as cosine similarity [62].

The image and QSD are heterogeneous modalities. Therefore, we adopt the cosine similarity to measure the relative difference, and design an attentive pooling for multimodal quality alignment. Specifically, we find the shared quality information by learning the attentive distribution, and the aligned visual feature $\hat{\mathbf{F}}_{img}$ is formulated by:

$$\begin{cases} \hat{\mathbf{F}}_{img} = \mathcal{N}_{ln} \left(\mathcal{F}_{fw} \left(\ddot{\mathbf{F}}_{img}; \theta_{fw} \right) + \ddot{\mathbf{F}}_{img} \right); \\ \ddot{\mathbf{F}}_{img} = \mathcal{N}_{ln} \left(\mathcal{F}_{ms} \left(\mathbf{F}_{img}; \theta_k, \theta_q, \theta_v, \sigma \right) + \mathbf{F}_{img} \right) \end{cases}, \quad (7)$$

where \mathcal{F}_{fw} denotes a feed-forward layer with the parameter weights θ_{fw} , and \mathcal{N}_{ln} denotes the layer normalization. \mathcal{F}_{ms} denotes a 32-head self-attention pooling module. θ_k , θ_q , and θ_v denote the parameter weights of key, query and value embedding in \mathcal{F}_{ms} , and we set the projection dimension size as 2048. $\sigma(\cdot) = \text{Softmax}(\cdot / \sqrt{d})$ denotes a scaled softmax used in \mathcal{F}_{ms} . Eq. (7) indicates that the feature elements related to shared quality information will obtain higher attention values than others. Therefore, the aligned quality information will be a weighted combination of all feature elements.

The aligned QSD feature $\hat{\mathbf{F}}_{aud}$ can be formulated as:

$$\begin{cases} \hat{\mathbf{F}}_{aud} = \mathcal{N}_{ln} \left(\mathcal{F}_{fw} \left(\ddot{\mathbf{F}}_{aud}; \theta'_{fw} \right) + \ddot{\mathbf{F}}_{aud} \right); \\ \ddot{\mathbf{F}}_{aud} = \mathcal{N}_{ln} \left(\mathcal{F}_{ms} \left(\mathbf{F}_{aud}; \theta'_k, \theta'_q, \theta'_v, \sigma \right) + \mathbf{F}_{aud} \right) \end{cases}, \quad (8)$$

where θ'_k , θ'_q , and θ'_v denote the embedding parameter weights. θ'_{fw} denotes the feed-forward parameter weights for the QSD feature alignment.

Finally, the relative difference between two aligned 1024-wide $\hat{\mathbf{F}}_{img}$ and $\hat{\mathbf{F}}_{aud}$ is measured by cosine similarity, which can be formulated by

$$\mathcal{D}_{cos} \left(\hat{\mathbf{F}}_{img}, \hat{\mathbf{F}}_{aud} \right) = \frac{\hat{\mathbf{F}}_{img} \odot \hat{\mathbf{F}}_{aud}}{\| \hat{\mathbf{F}}_{img} \| \| \hat{\mathbf{F}}_{aud} \|}, \quad (9)$$

where \odot represents an inner product operation and $\| \cdot \|$ represents the Euclidean distance. $\mathcal{D}_{cos} \left(\hat{\mathbf{F}}_{img}, \hat{\mathbf{F}}_{aud} \right)$ will be used as the metric of multimodal self-supervision learning in Sec. 4.4.1.

Algorithm 1 Proposed BMQA scheme on the image-audio case, $\text{BMQA}^{image-audio}$.

Input:

Image sample X_{img} and audio clip X_{aud} .

Output:

Quality score s_{pred} .

- 1: Extract the latent text feature by $\mathcal{F}_{aud2qsd}$ in Eq. (2);
 - 2: Extract the QSD feature representation \mathbf{F}_{aud} in Eqs. (3)-(4);
 - 3: Extract the image feature \mathbf{F}_{img} in Eqs. (5)-(6);
 - 4: Obtain the aligned image feature $\hat{\mathbf{F}}_{img}$ in Eq. (7) and the aligned QSD feature in $\hat{\mathbf{F}}_{aud}$ in Eq. (8), respectively;
 - 5: Fuse $\hat{\mathbf{F}}_{img}$ and $\hat{\mathbf{F}}_{aud}$, and predict the objective quality score s_{pred} in Eq. (10).
-

4.3 Multimodal Quality Fusion

In BMQA, the main benefit of using two modalities is that the image quality can be described from different perspectives [16]. For example, quality scores can be obtained directly from subjective tests, which can be reflected by the QSD information. Another representative case is that verbal descriptions of highly similar scenes are often unavoidably similar, while spatial details from images will help fine-grained scoring decisions. These two scenarios show that the fusion of QSD and image quality representations helps to predict more accurate quality scores. Therefore, multimodal quality fusion [3] is adopted to integrate two modalities to predict the quality score in our BMQA.

To preserve quality information as much as possible, we integrate two heterogenous modality features via a concatenation operation. Next, we employ a 2048-wide linear probe [62], \mathcal{F}_{fuse} , to fuse and forecast a final quality score s_{pred} , which can be formulated as:

$$s_{pred} = \mathcal{F}_{fuse} \left[\text{concat} \left(\hat{\mathbf{F}}_{img}, \hat{\mathbf{F}}_{aud} \right); \theta_{fuse} \right], \quad (10)$$

where $\text{concat}(\cdot)$ denotes the concatenation operation, and θ_{fuse} denotes the parameter weights of \mathcal{F}_{fuse} .

4.4 Deep Multimodal Learning Mechanism

BMQA learns to predict image quality scores by exploiting the supplementary quality descriptions from cross-modal features. In this section, we describe our multimodal learning mechanism, including multimodal self-supervision and supervision.

4.4.1 Multimodal Quality Self-Supervision

Existing methods suggest that image quality status can be latently learned without subjective scores [59]. In other words, quality status can be captured and learned from distorted image samples. The multimodal information is expected to enhance deep model learning, since the supplementary information from two modalities will contribute to the quality representation learning. For example, people prefer ‘bright’ images compared with ‘dark’ images.

The shared quality information of two modalities from the same instance provides supplementary information. Therefore, we learn an embedding space, in which feature pairs from the same instance gets close while these from different instances get far away from each other. Specifically, given the i -th $\hat{\mathbf{F}}_{img(i)}$ in the current training batch B , the

pair-wise probability $\mathcal{P}(\cdot)$ is calculated to find the matched j -th QSD feature as:

$$\mathcal{P}_{img}(i, j) = \frac{\exp\left(\mathcal{D}_{cos}\left(\hat{\mathbf{F}}_{img(i)}, \hat{\mathbf{F}}_{aud(j)}\right)/\tau\right)}{\sum_{k \in B} \exp\left(\mathcal{D}_{cos}\left(\hat{\mathbf{F}}_{img(i)}, \hat{\mathbf{F}}_{aud(k)}\right)/\tau\right)}, \quad (11)$$

where τ denotes the temperature parameter that controls the degree of distribution concentration.

At the same time, the pair-wise probability to find the matched i -th image feature for the j -th QSD feature $\hat{\mathbf{F}}_{aud(j)}$ will be:

$$\mathcal{P}_{aud}(j, i) = \frac{\exp\left(\mathcal{D}_{cos}\left(\hat{\mathbf{F}}_{img(i)}, \hat{\mathbf{F}}_{aud(j)}\right)/\tau\right)}{\sum_{k \in B} \exp\left(\mathcal{D}_{cos}\left(\hat{\mathbf{F}}_{img(k)}, \hat{\mathbf{F}}_{aud(j)}\right)/\tau\right)}. \quad (12)$$

Therefore, multimodal quality alignment needs to satisfy that the pair-wise probability between each paired $\hat{\mathbf{F}}_{img}$ and $\hat{\mathbf{F}}_{aud}$ gets the maximum value as: $\mathcal{P}_{img}(i, i) = \max_{k \in B} \mathcal{P}_{img}(i, k)$ and $\mathcal{P}_{aud}(j, j) = \max_{k \in B} \mathcal{P}_{aud}(j, k)$. The overall learning goal is to maximize the joint probability, which is equivalent to minimizing the negative log-likelihood:

$$\min_{\theta_{img}, \theta_{fw}, \theta_k, \theta_q, \theta_v; \theta_{aud}, \theta'_{fw}, \theta'_k, \theta'_q, \theta'_v} -\log \left[\sum_{i \in B} \mathcal{P}_{img}(i, i) + \sum_{j \in B} \mathcal{P}_{aud}(j, j) \right]. \quad (13)$$

4.4.2 Multimodal Quality Supervision

To obtain a better fusion model \mathcal{F}_{fuse} in Eq. (10), we adopt the mean square error $\|\cdot\|_2^2$ to measure the difference between the predicted score s_{pred} and the ground-truth subjective score s_{gt} . Our goal is to predict a score as close as possible to the subjective MOS, and the optimization problem can be formulated as:

$$\min_{\theta_{fuse}} \left\| \mathcal{F}_{fuse} \left[\text{concat} \left(\hat{\mathbf{F}}_{img}, \hat{\mathbf{F}}_{aud} \right); \theta_{fuse} \right] - s_{gt} \right\|_2^2. \quad (14)$$

4.5 Overall Algorithm

Given a pair of image data X_{img} and audio data X_{aud} , we obtain the image feature \mathbf{F}_{img} by Eq. (5) and the QSD feature \mathbf{F}_{aud} by Eq. (4). After that, BMQA learns to output the aligned features $\hat{\mathbf{F}}_{img}$ and $\hat{\mathbf{F}}_{aud}$ by Eq. (7) and Eq. (8), respectively. The learning of the quality feature alignment will be achieved by a multimodal self-supervision. Finally, BMQA learns to fuse the aligned feature by Eq. (10), where the learning of the quality feature fusion will be achieved by a multimodal supervision.

The detailed algorithm of our BMQA is summarized in **Algorithm 1**.

5 VALIDATIONS AND DISCUSSIONS

In this section, extensive experiments are conducted on two latest benchmark low-light image databases. Specifically, we verify the effectiveness of our BMQA on the image-audio database MLIQ (*i.e.*, $\text{BMQA}^{image-audio}$) and the image-only database Dark-4K (*i.e.*, $\text{BMQA}^{image-only}$), respectively. Besides, we demonstrate the comparison results with 25 competitive methods, including 8 hand-crafted BIQAs and 17 deep-learned BIQAs. We provide the detailed descriptions of the experimental validation, analysis, and discussion as follows.

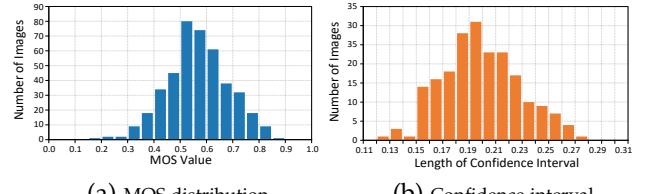


Fig. 9. A statistical analysis of the proposed unimodal Dark-4K database: (a) histogram of MOS values and (b) distribution of confidence interval.

5.1 Experimental Protocols

5.1.1 Benchmark Database

In the experiments, three databases are used including a pre-training database MS-COCO, a new multimodal image quality database MLIQ, and a new low-light database Dark-4K containing only a single image modality.

MS-COCO. We pre-train our BMQA model on the MS-COCO caption database [74], which contains 82783 image instances labeled with captions. The sentence of each caption contains at least 8 words. The partitions of MS-COCO strictly follow the officially specified setting.

MLIQ. We train our BMQA model on the proposed MLIQ dataset, containing 3600 pairs of low-light image and audio clip. The details can be found in Sec. 3. In the experiments, MLIQ is randomly divided by 8:1:1 according to the shooting scene. Specifically, the training set contains 576 scenes and 2880 samples, the validating set contains 72 scenes and 360 samples, and the testing set also contains 72 scenes and 360 samples.

Dark-4K. To validate the cross-dataset performance, we further establish a new ultra-high-definition (UHD) low-light database for the cross-dataset validation, namely Dark-4K. The original images of Dark-4K are collected from [10]. Dark-4K consists of 424 raw low-light images, which are captured by two consumer electronics: *Sony α7S-II* and *Fujifilm X-T2*. These two cameras have different imaging hardware: *Sony* has a full-frame Bayer sensor, and *Fujifilm* has an APS-C X-Trans sensor. Dark-4K supports the quality assessment of low-light images produced by different filter arrays.

The subjective experiments on Dark-4K maintain the same settings as described in Sec. 3.1.1. The histogram of labeled MOS results and the 95% confidence intervals for the subjective ratings are shown in Fig. 9 (a) and Fig. 9 (b), respectively.

5.1.2 Evaluation Metric

In visual quality assessment, Pearson Linear Correlation (PLCC), Spearman Rank Correlation Coefficient (SRCC), and Root Mean Square Prediction Error (RMSE) are three commonly-used evaluation metrics [72], [82], [103]. PLCC is used to measure the linear relationship between objective predictions and subjective scores, SRCC reflects the monotonicity of predictions, and RMSE measures the accuracy of predictions. For a promising method, PLCC and SRCC are close to 1, while RMSE is close to 0.

5.1.3 Training Detail

All experiments have been carried out on a computing server with *Intel(R) Xeon(R) Gold 6226R CPU@2.90GHz*, 38GB RAM, and *NVIDIA A100-PCIE GPU@40GB×6*.

TABLE 1
Overall performance comparison between 15 \mathcal{F}_{img} and \mathcal{F}_{qsd} combinations on the MLIQ database.

\mathcal{F}_{img}	\mathcal{F}_{qsd}	PLCC \uparrow	SRCC \uparrow	RMSE \downarrow
VGG-19	BoW	0.8538	0.8532	0.0885
	RNN	0.8670	0.8636	0.0869
	TransF	0.8731	0.8685	0.0847
RN-50	BoW	0.8743	0.8750	0.0845
	RNN	0.8859	0.8851	0.0824
	TransF	0.8989	0.8877	0.0830
EN-B4	BoW	0.8760	0.8719	0.0834
	RNN	0.8898	0.8886	0.0810
	TransF	0.8987	0.8922	0.0825
ViT-B32	BoW	0.8802	0.8774	0.0827
	RNN	0.8879	0.8871	0.0822
	TransF	0.9089	0.9040	0.0816
CNXT-B	BoW	0.8822	0.8844	0.0832
	RNN	0.8970	0.8984	0.0817
	TransF	0.9121	0.9065	0.0802

Next, we report the setting of each training stage, such as 1) pre-training, 2) self-supervised training, and 3) supervised training. Note that *Adam* is used as our optimization solver in *Python Toolbox PyTorch*.

Self-supervised pre-training (Stage_{PT}). We first pre-train the image feature extractor \mathcal{F}_{img} as well as the QSD feature extractor \mathcal{F}_{qsd} on the MS-COCO database. We take 768 samples as a batch (*i.e.*, 128×6) and pre-train our models for 50 epochs. We set the initial learning rate to 1e-3 and decay it by a *coseine* schedule.

Self-supervised training (Stage_{SS}). We train our BMQA model using the pairs of low-light image and the corresponding audio clip from the proposed MLIQ database. At this time, we reduce the batch size to 256 and set the total training epochs to 20, where the initial learning rate is fixed at 4e-5. In the experiments, we find that neither undertrained (*e.g.*, a loss greater than 1.0) nor overtrained (*e.g.*, a loss less than 0.6) models can achieve better performance. Empirically, we randomly pick the final model with the loss around 0.8.

Supervised training (Stage_{ST}). For VGG, ResNet and EfficientNet, we set the initial learning rate to 8e-5 and reduce the learning rate by a factor of 0.95 at the 150-th and 250-th epochs. For ViT and ConvNeXT, we set a smaller initial learning rate to 6.4e-5 and also reduce the learning rate by a factor of 0.95 at the 150-th and 250-th epochs. Finally, we keep the batch size as 16 and train the quality score predictor for 300 epochs.

5.2 Feature Representation Validation

To verify the effectiveness of two heterogenous modality feature extractors, we have conducted the experiments on several representative \mathcal{F}_{qsd} and \mathcal{F}_{img} models. Specifically, we take VGG-19, ResNet-50 (*denoted as* RN-50), EfficientNet (*denoted as* EN-B4), Vision-in-Transformer (*denoted as* ViT-B32), and ConvNeXT (*denoted as* CNXT-B) as the image feature encoders, respectively. In addition, we take Bag-of-Word (*denoted as* BoW), recurrent neural network (*denoted as* RNN), and Transformer (*denoted as* TransF) as the text feature encoders, respectively.

Considering that the prediction accuracy of BMQA is mainly related to the feature representation, we first explore the impact of different network backbones of \mathcal{F}_{img} ,

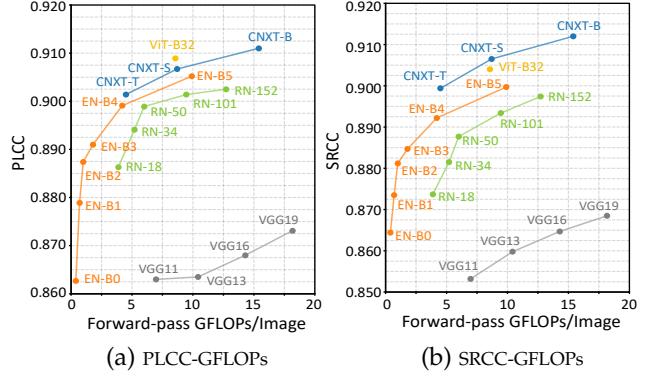


Fig. 10. Performance comparison of different image feature extractors \mathcal{F}_{img} : (a) PLCC-GFLOPs curves of different network variants and (b) SRCC-GFLOPs curves of different network variants.

including VGG19, RN-50, EN-B4, ViT-B32, and CNXT-B. Meanwhile, we employ BoW, RNN, and TransF as \mathcal{F}_{qsd} . Table 1 provides the overall comparison results of 15 \mathcal{F}_{qsd} and \mathcal{F}_{img} combinations. As seen, all 15 BMQA variants achieve promising performance, which verify the excellent robustness capability of our multimodal paradigm. Besides, the related combinations achieve higher prediction accuracy with the use of more powerful feature encoders.

In addition, considering that image is the main modality in the BIQA task, we further explore the impact of different network variants of \mathcal{F}_{img} . Fig. 10 provides the PLCC and SRCC curves of five \mathcal{F}_{img} models. As seen, we can draw some interesting conclusions: 1) Different variants of the same network architecture have a little effect on the final forecasting results. For example, the replacement of the ResNet network (from the heaviest RN-152 to the lightest RN-18) only results in the PLCC decreasing from 0.8863 to 0.9025 and the SRCC decreasing from 0.8737 to 0.8974; 2) The lightweight models have also achieved competitive performance. For example, EN-B0 achieves a PLCC score of 0.8627 and a SRCC score of 0.8532 with only 0.39G FLOPs. This suggests that our BMQA framework can maintain excellent performance for lightweight models.

5.3 Ablation Study

As mentioned above, BMQA is based on two modalities and is constructed through three training phases, including pre-training, multimodal self-supervision, and multimodal supervision. In view of this, the ablation experiments are conducted to verify the effectiveness of two different modalities for different learning stages. The detailed results are provided in Table 2.

Multimodal Self-supervision. To demonstrate the roles of two modalities, we first verify the performance of a single modality in Method (a) and Method (b), where “-” means the related modality or training is not used in Table 2. In other words, BMQA independently adopts RN-50 as the image encoder \mathcal{F}_{img} or TransF as the QSD encoder \mathcal{F}_{qsd} . As seen, the result of Method (a) is better than that of Method (b), which confirms that the image data is the main modality in the visual quality assessment task.

To verify the effect of Stage_{SS}, we also design Method (c) with the self-supervised learning based on Method (a), which is carried out for the feature alignment between TransF (\mathcal{F}_{qsd}) and RN-50 (\mathcal{F}_{img}) as described in Sec. 4.4.1.

TABLE 2
Ablation study of the contribution of each training stage:
 self-supervised pre-training Stage_{PT}, self-supervised training Stage_{SS},
 and supervised training Stage_{ST}.

Method	Stage _{PT}	Stage _{SS}		Stage _{ST}		PLCC↑	SRCC↑	RMSE↓
		\mathcal{F}_{img}	\mathcal{F}_{qsd}	\mathcal{F}_{img}	\mathcal{F}_{qsd}			
(a)	-	-	-	RN-50	-	0.7964	0.7934	0.1082
(b)	-	-	-	-	TransF	0.7504	0.7466	0.1130
(c)	-	RN-50	TransF	RN-50	-	0.8567	0.8556	0.0902
(d)	-	RN-50	TransF	-	TransF	0.8209	0.8183	0.0932
(e)	-	RN-50	TransF	RN-50	TransF	0.8614	0.8610	0.0896
(f)	CL	RN-50	TransF	RN-50	TransF	0.8707	0.8709	0.0868
(g)	RE	RN-50	TransF	RN-50	TransF	0.8753	0.8721	0.0871
(h)	FM	RN-50	TransF	RN-50	TransF	0.8989	0.8877	0.0830

As seen, Method (c) obtains a better result than Method (a), and it shows that the introduction of multimodal self-supervision helps the deep model training. The same phenomenon can be observed between Methods (d) and (b).

Multimodal Supervision. To verify the effect of Stage_{ST}, we also conduct the experiments with Method (e), which employs RN-50 as \mathcal{F}_{img} and TransF as \mathcal{F}_{qsd} to perform an end-to-end multimodal quality supervision. As seen, Method (e) achieves a better result than Method (c) and Method (d), which indicates that the introduction of multimodal supervised learning improves the performance of BMQA.

Pre-train Strategy. To show the feature alignment performance between \mathcal{F}_{qsd} and \mathcal{F}_{img} , we go a step further by pre-training Method (e) on the MS-COCO database. Specifically, we have performed three different pre-training strategies: 1) Method (f) verifies the pre-training effect on the classification task (*denoted as CL*), where RN-50 is pre-trained to distinguish categories of \mathcal{F}_{img} and TransF is pre-trained to distinguish categories of \mathcal{F}_{qsd} . 2) Method (g) verifies the pre-training effect on the restoration task (*denoted as RE*), where RN-50 is pre-trained to restore the missing image content in \mathcal{F}_{img} and TransF is pre-trained to fill up the missing word information in \mathcal{F}_{qsd} . 3) Method (h) verifies the pre-training effect on the feature alignment between \mathcal{F}_{img} and \mathcal{F}_{qsd} , where RN-50 and TransF are pre-trained to align the image and QSD features. Experiments show that Method (h) obtains a better result than Methods (f) and (g), which indicates that introducing multimodal quality alignment in the pre-training improves the BIQA performance.

5.4 Overall Performance Comparison

To demonstrate the overall quality forecasting performance, we compare our BMQA with 25 representative BIQA methods on the MLIQ database. For fairness, all comparison methods use the same experimental settings as described in Sec. 5.1. According to the type of quality feature extraction, these BIQAs are divided into two categories: 1) hand-crafted and 2) deep-learned methods.

Hand-crafted BIQAs. Considering that low-light images on MLIQ are characterized by the hybrid multiple distortions, the comparison methods based on hand-crafted features are composed of four types:

- **GP:** general-propose BIQAs, including *Mittal2012TIP* [56], *Liu2014SPIC* [41], and *Zhang2015TIP* [102].
- **MD:** distortion-specific BIQAs for multiply distortion, including *Li2016SPL* [37].

- **CD:** distortion-specific BIQAs for contrast distortion, including *Gu2017TCB* [24].
- **LL:** distortion-specific BIQAs for low-light distortion, including *Xiang2020TMM* [88], *Wang2021ICME* [79], and *Wang2022TII* [78].

It is noted that the implementation of *Xiang2020TMM* [88] is not available, and we provide the results of our reimplementation. For a fair comparison, all the other results in Table 3 are obtained by running the released implementations from the corresponding authors.

Deep-learned BIQAs. Considering that low-light images on MLIQ are characterized by some specific distortions, the comparison methods based on deep-learned features can be divided into four categories:

- **AS:** allocation-based supervised BIQAs, including *Kang2014CVPR* [30], *Bosse2018TIP* [4], and *Ke2021ICCV* [31].
- **GS:** generation-based supervised BIQAs, including *Ying2020CVPR* [99].
- **MT:** multi-task based supervised BIQAs, including *Kim2019TNNLS* [33], *Yan2019TMM* [94], *Zhang2020TCSVT* [104].
- **MO:** multi-objective based supervised BIQAs, including *Wu2020TIP* [86], *Li2020ACMMM* [36], *Su2020CVPR* [70], *Zhu2020CVPR* [111].
- **MC:** multi-objective based supervised BIQAs with crossdataset-learned constraint, including *Ma2021ACMMM* [48], *Zhang2021TIP* [105], and *Zhang2022TPAMI* [103].
- **MU:** metric-based unsupervised BIQAs, including *Wang2022ACMMM* [80] and *Madhusudana2022TIP* [49].
- **DU:** domain-based unsupervised BIQAs, including *Liu2019TPAMI* [43].

Performance Comparison. Table 3 tabulates the overall results of 25 representative BIQA methods. In addition, we provide the results of three BMQA variants, including BMQA_{RN-50+TransF}^{image-audio}, BMQA_{ViT-B32+TransF}^{image-audio} and BMQA_{CNXT-B+TransF}^{image-audio}, which take TransF as \mathcal{F}_{qsd} and take RN-50, ViT-B32, and CNXT-B as \mathcal{F}_{img} , respectively.

From Table 3, it is observed that the average prediction accuracy of three BMQAs significantly outperforms the other 25 competitive BIQA methods. The best BMQA_{CNXT-B+TransF}^{image-audio} obtains a PLCC score of 0.9121, a SRCC score of 0.9065, and a RMSE score of 0.0802, which is very promising in the BIQA task with authentic distortions. Furthermore, the experimental results also show that there is no significant performance difference for our BMQA from Device-I to Device-V, and the entire database, suggesting that different shooting devices have a slight side-effect in multimodal learning.

5.5 Further Discussion

In this section, we mainly discuss the applicability and generalization of our BMQA_{image-only} method in the current mainstream of single-image modality scenario.

5.5.1 Motivation

From Sec. 5.2 to 5.4, we have validated the effectiveness of our BMQA_{image-audio} in terms of feature representation,

TABLE 3

Performance comparison of the proposed BMQA method and 25 state-of-the-art BIQAs on the MLIQ database. The best results of the hand-crafted and deep-learned BIQAs are highlighted in **bold** for different devices, and the best results of our BMQA are highlighted in underline.

Method Type	Device-I (Nikon D5300)	Device-II (iPhone 8plus)	Device-III (iPad mini2)	Device-IV (Canon EOS)	Device-V (Huawei Mate)	Entire Database
Hand-crafted BIQAs Type	PLCC↑ SRCC↑ RMSE↓	PLCC↑ SRCC↑ RMSE↓	PLCC↑ SRCC↑ RMSE↓	PLCC↑ SRCC↑ RMSE↓	PLCC↑ SRCC↑ RMSE↓	PLCC↑ SRCC↑ RMSE↓
Mittal2012TIP [56]	GP 0.7833 0.7657 0.1073	0.7865 0.7695 0.0906	0.8005 0.7706 0.0802	0.7273 0.7136 0.1247	0.7709 0.7500 0.1100	0.7797 0.7716 0.1095
Liu2014SPIC [41]	GP 0.7848 0.7666 0.1072	0.7894 0.7710 0.0904	0.7777 0.7593 0.0904	0.7301 0.7178 0.1265	0.7678 0.7454 0.1109	0.7779 0.7689 0.1101
Zhang2015TIP [102]	GP 0.6537 0.6704 0.1303	0.6888 0.6747 0.1069	0.7883 0.8328 0.0837	0.5836 0.5832 0.1496	0.6736 0.6560 0.1286	0.6186 0.6346 0.1379
Li2016SPL [37]	MD 0.8267 0.8242 0.0961	0.7626 0.7335 0.0936	0.7935 0.7651 0.0821	0.8442 0.8338 0.0986	0.8767 0.8676 0.0824	0.8317 0.8273 0.0972
Gu2017TCB [24]	CD 0.8486 0.8446 0.0911	0.7957 0.7918 0.0893	0.7845 0.7878 0.0844	0.7721 0.7764 0.1171	0.7726 0.7658 0.1105	0.7726 0.7658 0.1105
Xiang2020TMM [88]	LL 0.8265 0.8221 0.1003	0.8642 0.8530 0.0885	0.8105 0.8004 0.0871	0.8323 0.8146 0.0730	0.8789 0.8642 0.0881	0.8063 0.7950 0.1011
Wang2021ICME [79]	LL 0.8619 0.8539 0.0868	0.8304 0.8179 0.0798	0.8002 0.7759 0.0805	0.8584 0.8538 0.0942	0.8386 0.8234 0.0954	0.8365 0.8350 0.0969
Wang2022TI [78]	LL 0.8647 0.8571 0.0845	0.8401 0.8373 0.0815	0.8336 0.8459 0.0713	0.9085 0.8892 0.0789	0.8577 0.8484 0.0936	0.8578 0.8488 0.0904
Deep-learned BIQAs Type	PLCC↑ SRCC↑ RMSE↓	PLCC↑ SRCC↑ RMSE↓	PLCC↑ SRCC↑ RMSE↓	PLCC↑ SRCC↑ RMSE↓	PLCC↑ SRCC↑ RMSE↓	PLCC↑ SRCC↑ RMSE↓
Kang2014CVPR [30]	AS 0.8103 0.8197 0.1580	0.7954 0.7987 0.1537	0.6822 0.6992 0.1792	0.8301 0.8629 0.1681	0.8783 0.8874 0.1575	0.8095 0.8167 0.1604
Bosse2018TIP [4]	AS 0.9077 0.9154 0.0918	0.7395 0.7198 0.0979	0.6824 0.6305 0.1281	0.8695 0.8596 0.1198	0.8516 0.8615 0.1164	0.8182 0.8198 0.1070
Ke2021ICCV [31]	AS 0.8614 0.8464 0.0884	0.8343 0.8254 0.0822	0.8297 0.8208 0.0738	0.8956 0.8835 0.0800	0.8362 0.8312 0.0944	0.8504 0.8487 0.0923
Ying2020CVPR [99]	GS 0.8458 0.8427 0.0926	0.8561 0.8468 0.0893	0.8117 0.7953 0.0854	0.8628 0.8138 0.0700	0.8842 0.8616 0.0846	0.8450 0.8455 0.0921
Kim2019TNLNS [33]	MT 0.9096 0.9166 0.0863	0.7424 0.7213 0.0963	0.6828 0.6305 0.1333	0.8713 0.8609 0.1213	0.8515 0.8633 0.1180	0.8217 0.8234 0.1063
Yan2019TMM [94]	MT 0.8244 0.8276 0.1631	0.8030 0.8105 0.1575	0.7156 0.6992 0.1755	0.8690 0.8775 0.1700	0.8398 0.8535 0.1624	0.8256 0.8370 0.1639
Zhang2020TCVTS [104]	MT 0.8720 0.8845 0.1075	0.7118 0.7141 0.1035	0.6681 0.5929 0.1250	0.8423 0.8340 0.1163	0.7631 0.7558 0.1213	0.8241 0.8261 0.1035
Wu2020TIP [86]	MO 0.8713 0.8786 0.1249	0.7132 0.7188 0.1163	0.8516 0.8277 0.0721	0.9265 0.9356 0.0976	0.8537 0.8498 0.0989	0.8299 0.8371 0.1099
Li2020ACMMM [36]	MO 0.8561 0.8514 0.0877	0.7981 0.7894 0.0882	0.8074 0.8275 0.0706	0.8866 0.8740 0.0869	0.8174 0.8028 0.0995	0.8236 0.8223 0.0988
Su2020CVPR [70]	MO 0.8645 0.8581 0.0843	0.7933 0.7898 0.0791	0.8171 0.8226 0.0664	0.9303 0.9250 0.0677	0.8401 0.8427 0.0893	0.8520 0.8510 0.0882
Zhu2020CVPR [111]	MO 0.9248 0.9311 0.0721	0.7709 0.7626 0.0922	0.6819 0.6283 0.1211	0.8890 0.8814 0.1096	0.8513 0.8603 0.1031	0.8519 0.8547 0.0948
Ma2021ACMMM [48]	MC 0.8594 0.8623 0.0871	0.7971 0.7927 0.0877	0.8049 0.8017 0.0771	0.8631 0.8592 0.0903	0.8092 0.7997 0.0987	0.8271 0.8275 0.0982
Zhang2021TIP [105]	MC 0.8848 0.8838 0.0880	0.8487 0.8466 0.0832	0.8479 0.8051 0.0708	0.8476 0.8482 0.1052	0.8147 0.8135 0.1008	0.8485 0.8605 0.0929
Zhang2022TPAMI [103]	MC 0.9277 0.9320 0.0713	0.7808 0.7712 0.0911	0.6811 0.6403 0.1117	0.8938 0.8870 0.1037	0.8908 0.8977 0.0975	0.8596 0.8615 0.0910
Liu2019PTPAMI [43]	DU 0.8507 0.8451 0.1105	0.8185 0.8100 0.0979	0.8096 0.8269 0.0644	0.8300 0.8121 0.1095	0.7593 0.7346 0.1206	0.8103 0.8140 0.1081
Wang2022ACM [80]	MU 0.8657 0.8608 0.0870	0.8210 0.8206 0.0854	0.8265 0.8122 0.0747	0.8764 0.8665 0.0862	0.8311 0.8215 0.0994	0.8383 0.8373 0.0973
Madhusudana2022TIP [49]	MU 0.9179 0.9254 0.0784	0.7708 0.7656 0.0935	0.7152 0.6569 0.1099	0.8908 0.8842 0.1042	0.8474 0.8579 0.1017	0.8495 0.8531 0.0943
BMQA ^{Image-audio} _{RN50+TransF}	MU 0.9348 0.9187 0.0695	0.9064 0.8813 0.0686	0.9075 0.8712 0.0638	0.9364 0.9140 0.0726	0.9011 0.8860 0.0816	0.8989 0.8877 0.0830
BMQA ^{Image-audio} _{IT-B32+TransF}	MU 0.9320 0.9264 0.0715	0.9131 0.9042 0.0664	0.9108 0.9067 0.0598	0.9488 0.9282 0.0737	0.9041 0.8913 0.0828	0.9089 0.9040 0.0816
BMQA ^{Image-audio} _{CNXT-B+TransF}	MU 0.9404 0.9310 0.0692	0.9128 0.9022 0.0653	0.8920 0.8744 0.0607	0.9464 0.9358 0.0717	0.8959 0.8850 0.0808	0.9121 0.9065 0.0802

ablation study, and overall performance. Nevertheless, the proposed BMQA has to face the challenge due to the absence of QSD-based audio clips.

Frankly speaking, it is not easy to require QSD-based audio clips for existing low-light BIQA scenarios. In other words, the question is how to use our BMQA when the QSD feature is absent. To address this challenge, we further propose a feasible scheme and validate it on another independent low-light database, Dark-4K, which contains only image samples and their MOS labels.

5.5.2 Additional Modality Generation

Considering that the QSD feature can be mainly represented by the related language information in the BIQA task, we believe that $\mathcal{F}_{aud2qsd}$ can also be replaced by a recent image caption model: ‘show and tell’ (SAT) [90]. Therefore, when the QSD feature is unavailable, BMQA employs an image captioning model to generate the feature representation of $\mathcal{F}_{aud2qsd}$.

To the best of our survey, few existing image captioning models are specially trained for the BIQA task, and hence the generated caption is less relevant to the visual quality experience, which may not meet the QSD principles described in Sec. 3.1.2. Therefore, we train a special QSD-based captioning model $\mathcal{F}_{img2qsd}$ based on the MLIQ database (e.g., 8:1:1 partition).

Specifically, each image sample is fed into $\mathcal{F}_{img2qsd}$, and the verbal description of its audio clip is used as the training label. The learning goal is to maximize the joint pair-wise probability of the predicted caption and the annotated cap-



Fig. 11. Examples of the generated QSD features on our Dark-4K database. Each example contains an original image sample (left), an attention map of the QSD feature related to quality cognition (right), a predicted caption result obtained by a pre-trained SAT caption model ($SAT_{MS-COCO}$) on the MS-COCO database, and a low-light QSD-based caption result of SAT_{MLIQ} trained on the MLIQ database.

tion, or equivalently to minimize the log-likelihood over the training set as:

$$\min \mathbb{E} \left\{ -\log \sum_{1 \leq k \leq N} \mathcal{P} [\mathcal{F}_{img2qsd}(X_{img}; \theta_{img2qsd}), x_k] \right\}, \quad (15)$$

where $\theta_{img2qsd}$ represents the model weights of $\mathcal{F}_{img2qsd}$.

Therefore, given an image X_{img} , the QSD feature is generated by

$$< x'_1, \dots, x'_N > = \mathcal{F}_{img2qsd}(X_{img}; \theta_{img2qsd}), \quad (16)$$

where $< x'_1, \dots, x'_N >$ denotes the predicted QSD text.

To verify the feasibility of $\mathcal{F}_{img2qsd}$, we directly adopt the SAT model as the backbone of $\mathcal{F}_{img2qsd}$. In the experiments, we employ a pre-trained model provided by [90] (denoted as $SAT_{MS-COCO}$), and finetune a special QSD-based captioning model on the MLIQ database for 100 epochs (denoted as SAT_{MLIQ}). Finally, the TOP-5 accuracy of caption matching performance on the testing set reaches 89.74%, and the BLEU-4 score reaches 29.36%.

Performance comparison of the proposed BMQA and 17 deep-learned BIQAs on the Dark-4K database. The best results of the deep-learned BIQAs are highlighted in **bold**, and the best results of our BMQA are highlighted in underline.

Method Type		Entire Database		
Deep-learned BIQAs	Type	PLCC \uparrow	SRCC \uparrow	RMSE \downarrow
Kang2014CVPR [30]	AS	0.5434	0.5605	0.1307
Bosse2018TIP [4]	AS	0.6552	0.6442	0.1028
Ke2021ICCV [31]	AS	0.8630	<u>0.8382</u>	0.1195
Ying2020CVPR [99]	GS	0.6453	0.5802	0.0631
Kim2019TNNLS [33]	MT	0.5968	0.6030	0.1205
Yan2019TMM [94]	MT	0.5588	0.5610	0.1322
Zhang2020TCSV [104]	MT	0.5187	0.5581	0.1346
Wu2020TIP [86]	MO	0.6216	0.5877	0.1248
Li2020ACMMM [36]	MO	0.6138	0.6358	0.1107
Su2020CVPR [70]	MO	0.7914	0.7829	0.0755
Zhu2020CVPR [111]	MO	0.7791	0.7616	0.0931
Ma2021ACMMM [48]	MC	0.7713	0.7419	0.0810
Zhang2021TIP [105]	MC	0.8079	0.8131	0.0697
Zhang2022TPAMI [103]	MC	0.8538	0.8358	<u>0.0593</u>
Liu2019TPAMI [43]	DU	0.6319	0.6352	0.1002
Wang2022ACMMM [80]	MU	0.8526	0.8357	0.0687
Madhusudana2022TIP [49]	MU	0.8293	0.8240	0.0669
BMQA ^{image-only} RN+TransF	MU	0.9026	0.8874	0.0624
BMQA ^{image-only} ViT-B32+TransF	MU	0.9117	0.8989	<u>0.0593</u>
BMQA ^{image-only} CNXT-B+TransF	MU	0.9156	<u>0.9085</u>	0.0605

Fig. 11 shows some QSD-based caption results of SAT_{MLIQ} and SAT_{MS-COCO} on the Dark-4K database. As seen, SAT_{MLIQ} contains more QSD features related to the visual quality experience, such as brightness status, salient objects, and color status. To further explore the relationship between these keywords and the original images, we provide the attention maps corresponding to these keywords. As seen, the region of interest (ROI) is consistent with the QSD keywords. Therefore, we believe that a well-trained QSD-based captioning model is competent to generate the QSD modality provided in MLIQ.

5.5.3 Cross-dataset Validation on Image-only Case

To verify the effect of $\mathcal{F}_{img2qsd}$ in Eq. (16), we conduct the cross-database validation on the Dark-4K database. For a fair comparison, all BIQAs including our BMQA models have not been fine-tuned or retrained. In other words, all trained models are kept exactly the same as in Sec. 5.4. It is worth noting that since $\mathcal{F}_{aud2qsd}$ does not exist on the Dark-4K database, a specially-learned model $\mathcal{F}_{img2qsd}$ is used to replace $\mathcal{F}_{aud2qsd}$ in Algorithm 1. Apart from this, other settings are the same as in Sec. 5.1.3.

Table 4 provides the overall comparison results of all 17 deep-learned BIQAs in terms of PLCC, SRCC, and RMSE. As seen, our BMQAs achieve the state-of-the-art performance on the cross-dataset validation. The best BMQA^{image-only}
CNXT-B+TransF obtains a PLCC score of 0.9156, a SRCC score of 0.9085, and a RMSE score of 0.0605. Experimental results verify the applicability and generalization performance of our BMQA framework.

6 CONCLUSION AND FUTURE WORK

In this article, we have presented a new study on blind multimodal quality assessment (BMQA) of low-light images from both subjective and objective perspectives. Specifically, we establish the first multimodal quality assessment database for low-light images, where two quality-aware

principles are designed for multimodal benchmark construction. Moreover, we investigate the BMQA framework by exploring quality feature representation, alignment and fusion, and multimodal learning. Experimental results verify the effectiveness of our BMQA compared with state-of-the-art methods on two different low-light benchmark databases. We believe that this work can bring a new research perspective to image and video quality assessment.

In the future, BMQA can be further studied and verified in many aspects. First, we demonstrate how multimodal data is able to improve the performance of blind image quality assessment (BIQA) on the low-light application scenario, rather than how multimodal data collectively determines the visual quality experience. The latter is more challenging and thus requires deploying more sensors and fusing multi-sensor data.

Second, we design two quality semantic description (QSD) principles in the audio clip collection and only extract the text semantic feature in multimodal learning. However, audio information in actual scenes can be very complex and hence more audio attributes (*e.g.*, timbre, pitch, rhythm, *etc.*) need to be further investigated.

Third, we employ the multimodal self-supervision and supervision mechanisms to train our baseline model. The current design is simple and efficient on low-light images, but the interaction between different modalities needs to be further explored for the BMQA framework, which will help to extract more efficient feature representations.

Last but not least, we construct a multimodal low-light image database and present an effective no-reference quality indicator because it is more challenging and urgent. However, the feasibility of multimodality-driven quality assessment needs to be verified on more benchmark databases, including those for distortion-specific and general-purpose applications. Furthermore, its feasibility needs to be studied and verified on the full-reference task, as this quality assessment paradigm is still used in many real-world scenarios.

REFERENCES

- [1] Dario Amodei, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, et al. Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin. In *PMLR International Conference on Machine Learning (ICML)*, pages 173–182, 2016.
- [2] Yongqiang Bai, Zhongjie Zhu, Gangyi Jiang, and Huifang Sun. Blind quality assessment of screen content images via macro-micro modeling of tensor domain dictionary. *IEEE Transactions on Multimedia*, 23(1):4259–4271, 2021.
- [3] Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2):423–443, 2018.
- [4] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, 2018.
- [5] ITU. Methodologies for the subjective assessment of the quality of television images, document Recommendation ITU-R BT.500-14 (10/2019). *ITU, Geneva, Switzerland*, 2020.
- [6] Peibei Cao, Zhangyang Wang, and Kede Ma. Debiased subjective assessment of real-world image enhancement. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 711–721, 2021.

- [7] Yaqin Cao, Xiongkuo Min, Wei Sun, and Guangtao Zhai. Deep Neural Networks For Full-Reference And No-Reference Audio-Visual Quality Assessment. In *IEEE International Conference on Image Processing (ICIP)*, pages 1429–1433, 2021.
- [8] Bingzhi Chen, Qi Cao, Mixiao Hou, Zheng Zhang, Guangming Lu, and David Zhang. Multimodal Emotion Recognition With Temporal and Semantic Consistency. *IEEE Transactions on Audio, Speech, and Language Processing*, 29(1):3592–3603, 2022.
- [9] Baoliang Chen, Haoliang Li, Hongfei Fan, and Shiqi Wang. No-reference screen content image quality assessment with unsupervised domain adaptation. *IEEE Transactions on Image Processing*, 30(1):5463–5476, 2021.
- [10] Chen Chen, Qifeng Chen, Jia Xu, and Vladlen Koltun. Learning to see in the dark. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3291–3300, 2018.
- [11] Liang Chen, Jiawei Zhang, Jinshan Pan, Songnan Lin, Faming Fang, and Jimmy S Ren. Learning a non-blind deblurring network for night blurry images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10542–10550, 2021.
- [12] Pengfei Chen, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. Unsupervised curriculum domain adaptation for no-reference video quality assessment. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5178–5187, 2021.
- [13] Ziteng Cui, Guo-Jun Qi, Lin Gu, Shaodi Yu, Zenghui Zhang, and Tatsuya Harada. Multitask AET with orthogonal tangent regularity for dark object detection. In *IEEE International Conference on Computer Vision (ICCV)*, pages 2553–2562, 2021.
- [14] Xueqing Deng, Peng Wang, Xiaochen Lian, and Shawn Newsam. NightLab: A dual-level architecture with hardness detection for segmentation at night. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 16938–16948, 2022.
- [15] Yubin Deng, Chen Change Loy, and Xiaou Tang. Image aesthetic assessment: An experimental survey. *IEEE Signal Processing Magazine*, 34(4):80–106, 2017.
- [16] Sidney K D'mello and Jacqueline Kory. A review and meta-analysis of multimodal affect detection systems. *ACM Computing Surveys*, 47(3):1–36, 2015.
- [17] David Dov, Ronen Talmon, and Israel Cohen. Audio-visual voice activity detection using diffusion maps. *IEEE Transactions on Audio, Speech, and Language Processing*, 23(4):732–745, 2015.
- [18] Yuming Fang, Liping Huang, Jiebin Yan, Xuelin Liu, and Yang Liu. Perceptual quality assessment of omnidirectional images. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 580–588, 2022.
- [19] Yuming Fang, Jiebin Yan, Leida Li, Jinjian Wu, and Weisi Lin. No reference quality assessment for screen content images with both local and global feature representation. *IEEE Transactions on Image Processing*, 27(4):1600–1610, 2017.
- [20] Pedro Garcia Freitas, Welington YL Akamine, and Mylene CQ Farias. No-reference image quality assessment using orthogonal color planes patterns. *IEEE Transactions on Multimedia*, 20(12):3353–3360, 2018.
- [21] Ian E Gordon. *Theories of visual perception*. Psychology press, 2004.
- [22] Richard Langton Gregory. Perceptions as hypotheses. *Philosophical Transactions of the Royal Society of London. B, Biological Sciences*, 290(1038):181–197, 1980.
- [23] Jie Gu, Gaofeng Meng, Cheng Da, Shimeng Xiang, and Chunhong Pan. No-reference image quality assessment with reinforcement recursive list-wise ranking. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 8336–8343, 2019.
- [24] Ke Gu, Jun Zhou, Jun-Fei Qiao, Guangtao Zhai, Weisi Lin, and Alan Conrad Bovik. No-reference quality assessment of screen content pictures. *IEEE Transactions on Image Processing*, 26(8):4005–4018, 2017.
- [25] David S Hands. A basic multimedia quality model. *IEEE Transactions on Multimedia*, 6(6):806–816, 2004.
- [26] Alan Hanjalic and Li-Qun Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.
- [27] Mingkai Huang, Qiu Shen, Zhan Ma, Alan Conrad Bovik, Praful Gupta, Rongbing Zhou, and Xun Cao. Modeling the perceptual quality of immersive images rendered on head mounted displays: Resolution and compression. *IEEE Transactions on Image Processing*, 27(12):6039–6050, 2018.
- [28] Hao Jiang, Gangyi Jiang, Mei Yu, Yun Zhang, You Yang, Zongju Peng, Fen Chen, and Qingbo Zhang. Cubemap-based perception-driven blind quality assessment for 360-degree images. *IEEE Transactions on Image Processing*, 30(1):2364–2377, 2021.
- [29] Mingxing Jiang, Liquan Shen, Min Hu, Ping An, and Fuji Ren. Blind Quality Evaluator of Tone-Mapped HDR and Multi-Exposure Fused Images for Electronic Display. *IEEE Transactions on Consumer Electronics*, 67(4):350–362, 2021.
- [30] Le Kang, Peng Ye, Yi Li, and David Doermann. Convolutional neural networks for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1733–1740, 2014.
- [31] Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. MUSIQ: Multi-scale image quality transformer. In *IEEE International Conference on Computer Vision (ICCV)*, pages 5148–5157, 2021.
- [32] Jongyoo Kim and Sanghoon Lee. Fully deep blind image quality predictor. *IEEE Journal of Selected Topics in Signal Processing*, 11(1):206–220, 2017.
- [33] Jongyoo Kim, Anh-Duc Nguyen, and Sanghoon Lee. Deep CNN-based blind image quality predictor. *IEEE Transactions on Neural Networks and Learning Systems*, 30(1):11–24, 2019.
- [34] Debarati Kundu, Deepthi Ghadiyaram, Alan C Bovik, and Brian L Evans. No-reference quality assessment of tone-mapped HDR pictures. *IEEE Transactions on Image Processing*, 26(6):2957–2971, 2017.
- [35] Chongyi Li, Chunle Guo, Ling-Hao Han, Jun Jiang, Ming-Ming Cheng, Jinwei Gu, and Chen Change Loy. Low-light image and video enhancement using deep learning: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(12):9396–9416, 2022.
- [36] Dingquan Li, Tingting Jiang, and Ming Jiang. Norm-in-norm loss with faster convergence and better performance for image quality assessment. In *ACM International Conference on Multimedia (MM)*, pages 789–797, 2020.
- [37] Qiaohong Li, Weisi Lin, and Yuming Fang. No-reference quality assessment for multiply-distorted images in gradient domain. *IEEE Signal Processing Letters*, 23(4):541–545, 2016.
- [38] Qiaohong Li, Weisi Lin, Jingtao Xu, and Yuming Fang. Blind image quality assessment using statistical structural and luminance features. *IEEE Transactions on Multimedia*, 18(12):2457–2469, 2016.
- [39] Kwan-Yee Lin and Guanxiang Wang. Hallucinated-IQA: No-reference image quality assessment via adversarial learning. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 732–741, 2018.
- [40] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Anan Liu, Bin Wang, and Yongdong Zhang. Focus Your Attention: A Focal Attention for Multimodal Learning. *IEEE Transactions on Multimedia*, 24(1):103–115, 2022.
- [41] Lixiong Liu, Bao Liu, Hua Huang, and Alan Conrad Bovik. No-reference image quality assessment based on spatial and spectral entropies. *Elsevier Signal Processing: Image Communication*, 29(8):856–863, 2014.
- [42] Tsung-Jung Liu and Kuan-Hsien Liu. No-reference image quality assessment by wide-perceptual-domain scorer ensemble method. *IEEE Transactions on Image Processing*, 27(3):1138–1151, 2018.
- [43] Xialei Liu, Joost Van De Weijer, and Andrew D Bagdanov. Exploiting unlabeled data in CNNs by self-supervised learning to rank. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(8):1862–1878, 2019.
- [44] Yutao Liu, Ke Gu, Xiu Li, and Yongbing Zhang. Blind image quality assessment by natural scene statistics and perceptual characteristics. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 16(3):1–91, 2020.
- [45] Yuen Peng Loh and Chee Seng Chan. Getting to know low-light images with the exclusively dark dataset. *Elsevier Computer Vision and Image Understanding*, 178:30–42, 2019.
- [46] Kede Ma, Wentao Liu, Tongliang Liu, Zhou Wang, and Dacheng Tao. dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Transactions on Image Processing*, 26(8):3951–3964, 2017.
- [47] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing*, 27(3):1202–1213, 2018.
- [48] Rui Ma, Hanxiao Luo, Qingbo Wu, King Ngi Ngan, Hongliang Li, Fanman Meng, and Linfeng Xu. Remember and Reuse: Cross-Task Blind Image Quality Assessment via Relevance-aware Incremental Learning. In *ACM International Conference on Multimedia (MM)*, pages 5248–5256, 2021.
- [49] Pavan C Madhusudana, Neil Birkbeck, Yilin Wang, Balu Adsumilli, and Alan C Bovik. Image quality assessment using contrastive learning. *IEEE Transactions on Image Processing*, 31(1):4149–4161, 2022.
- [50] Saeed Mahmoudpour and Peter Schelkens. A Multi-Attribute Blind Quality Evaluator for Tone-Mapped Images. *IEEE Transactions on Multimedia*, 22(8):1939–1954, 2020.

- [51] Helard A Becerra Martinez and Mylene CQ Farias. Combining audio and video metrics to assess audio-visual quality. *Springer Multimedia Tools and Applications*, 77(18):23993–24012, 2018.
- [52] Harry McGurk and John MacDonald. Hearing lips and seeing voices. *Nature*, 264(5588):746–748, 1976.
- [53] Xiongkuo Min, Ke Gu, Guangtao Zhai, Xiaokang Yang, Wenjun Zhang, Patrick Le Callet, and Chang Wen Chen. Screen content quality assessment: Overview, benchmark, and beyond. *ACM Computing Surveys*, 54(9):1–36, 2021.
- [54] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Mylene CQ Farias, and Alan Conrad Bovik. Study of subjective and objective quality assessment of audio-visual signals. *IEEE Transactions on Image Processing*, 29(5588):6054–6068, 2020.
- [55] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang, and Xinping Guan. A multimodal saliency model for videos with high audio-visual correspondence. *IEEE Transactions on Image Processing*, 29:3805–3819, 2020.
- [56] Anish Mittal, Anush Krishna Moorthy, and Alan Conrad Bovik. No-reference image quality assessment in the spatial domain. *IEEE Transactions on Image Processing*, 21(12):4695–4708, 2012.
- [57] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a ‘Completely Blind’ Image Quality Analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2013.
- [58] Anush Krishna Moorthy and Alan Conrad Bovik. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Transactions on Image Processing*, 20(12):3350–3364, 2011.
- [59] Fu-Zhao Ou, Xingyu Chen, Ruixin Zhang, Yuge Huang, Shaoxin Li, Jilin Li, Yong Li, Luijuan Cao, and Yuan-Gen Wang. Sdd-fiq: Unsupervised face image quality assessment with similarity distribution distance. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 7670–7679, 2021.
- [60] Margaret H Pinson, Lucjan Janowski, Romuald Pépion, Quan Huynh-Thu, Christian Schmidmer, Phillip Corriveau, Audrey Younkin, Patrick Le Callet, Marcus Barkowsky, and William Ingram. The influence of subjects and environment on audiovisual subjective tests: An international study. *IEEE Journal of Selected Topics in Signal Processing*, 6(6):640–651, 2012.
- [61] Fatih Porikli, Al Bovik, Chris Plack, Ghassan AlRegib, Joyce Farrell, Patrick Le Callet, Quan Huynh-Thu, Sebastian Möller, and Stefan Winkler. Multimedia quality assessment [DSP Forum]. *IEEE Signal Processing Magazine*, 28(6):164–177, 2011.
- [62] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *PMLR International Conference on Machine Learning (ICML)*, pages 8748–8763, 2021.
- [63] Hongyu Ren, Diqi Chen, and Yizhou Wang. RAN4IQA: Restorative adversarial nets for no-reference image quality assessment. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 7308–7314, 2018.
- [64] Michele A Saad, Alan C Bovik, and Christophe Charrier. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Transactions on Image Processing*, 21(8):3339–3352, 2012.
- [65] Shaheer U. Saeed, Yunguan Fu, Vasilis Stavrinides, Zachary M.C. Baum, Qianye Yang, Mirabela Rusu, Richard E. Fan, Geoffrey A. Sonn, J. Alison Noble, Dean C. Barratt, and Yipeng Hu. Image quality assessment for machine learning tasks using meta-reinforcement learning. *Elsevier Medical Image Analysis*, 78(1):102427, 2022.
- [66] Ran Shi, Jing Ma, King Ngi Ngan, Jian Xiong, and Tong Qiao. Objective Object Segmentation Visual Quality Evaluation: Quality Measure and Pooling Method. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 18(3):1–19, 2022.
- [67] Guoli Song, Shuhui Wang, Qingming Huang, and Qi Tian. Harmonized multimodal learning with Gaussian process latent variable models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(3):858–872, 2019.
- [68] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3008–3021, 2020.
- [69] Russell L Storms and Michael J Zyda. Interactions in perceived quality of auditory-visual displays. *MIT Press Presence: Teleoperators & Virtual Environments*, 9(6):557–580, 2000.
- [70] Shaolin Su, Qingsen Yan, Yu Zhu, Cheng Zhang, Xin Ge, Jinqui Sun, and Yanning Zhang. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3667–3676, 2020.
- [71] Xiangjie Sui, Kede Ma, Yiru Yao, and Yuming Fang. Perceptual quality assessment of omnidirectional images as moving camera videos. *IEEE Transactions on Visualization and Computer Graphics*, 28(8):3022–3034, 2022.
- [72] Raoni FS Teixeira and Neucimar J Leite. A new framework for quality assessment of high-resolution fingerprint images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(10):1905–1917, 2016.
- [73] Yu Tian, Huanqiang Zeng, Junhui Hou, Jing Chen, Jianqing Zhu, and Kai-Kuang Ma. A light field image quality assessment model based on symmetry and depth features. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):2046–2050, 2021.
- [74] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: Lessons learned from the 2015 mscoco image captioning challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4):652–663, 2016.
- [75] Nicholas Wade and Mike Swanston. *Visual perception: An introduction*. Psychology Press, 2013.
- [76] Markus Waltl, Christian Timmerer, and Hermann Hellwagner. Improving the quality of multimedia experience through sensory effects. In *IEEE International Workshop on Quality of Multimedia Experience (QoMEX)*, pages 124–129, 2010.
- [77] Guotao Wang, Chenglizhao Chen, Deng-Ping Fan, Aimin Hao, and Hong Qin. From semantic categories to fixations: A novel weakly-supervised visual-auditory saliency detection approach. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 15119–15128, 2021.
- [78] Miaohui Wang, Yijing Huang, Jian Xiong, and Wuyuan Xie. Low-light Images In-the-wild: A Novel Visibility Perception-guided Blind Quality Indicator. *IEEE Transactions on Industrial Informatics*, 1(1):1–11, 2022.
- [79] Miaohui Wang, Yijing Huang, and Jialin Zhang. Blind Quality Assessment of Night-Time Images Via Weak Illumination Analysis. In *IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2021.
- [80] Miaohui Wang, Zhuowei Xu, Yuanhao Gong, and Wuyuan Xie. S-CCR: Super-Complete Comparative Representation for Low-Light Image Quality Inference In-the-wild. In *ACM International Conference on Multimedia (MM)*, pages 5219–5227, 2022.
- [81] Xuejin Wang, Feng Shao, Qiuping Jiang, Zhenqi Fu, Xiangchao Meng, Ke Gu, and Yo-Sung Ho. Combining Retargeting Quality and Depth Perception Measures for Quality Evaluation of Retargeted Stereopairs. *IEEE Transactions on Multimedia*, 24(1):2422–2434, 2022.
- [82] Zhihua Wang and Kede Ma. Active fine-tuning from gMAD examples improves blind image quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(1):1–14, 2021.
- [83] Xuekai Wei, Jing Li, Mingliang Zhou, and Xianmin Wang. Contrastive distortion-level learning-based no-reference image-quality assessment. *Wiley International Journal of Intelligent Systems*, 37(11):8730–8746, 2022.
- [84] Wenyng Wen, Kangkang Wei, Yuming Fang, and Yushu Zhang. Visual quality assessment for perceptually encrypted light field images. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(7):2522–2534, 2020.
- [85] Stefan Winkler and Christof Faller. Perceived audiovisual quality of low-bitrate multimedia content. *IEEE transactions on Multimedia*, 8(5):973–980, 2006.
- [86] Jinjian Wu, Jupo Ma, Fuhu Liang, Weisheng Dong, Guangming Shi, and Weisi Lin. End-to-end blind image quality prediction with cascaded deep neural network. *IEEE Transactions on Image Processing*, 29(1):7414–7426, 2020.
- [87] Qingbo Wu, Lei Wang, King Ngi Ngan, Hongliang Li, Fanman Meng, and Linfeng Xu. Subjective and objective de-raining quality assessment towards authentic rain image. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(11):3883–3897, 2020.
- [88] Tao Xiang, Ying Yang, and Shangwei Guo. Blind night-time image quality assessment: Subjective and objective approaches. *IEEE Transactions on Multimedia*, 22(5):1259–1272, 2019.
- [89] Jingtao Xu, Peng Ye, Qiaohong Li, Haiqing Du, Yong Liu, and David Doermann. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing*, 25(9):4444–4457, 2016.
- [90] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *PMLR International Conference on Machine Learning (ICML)*, pages 2048–2057, 2015.

- [91] Mai Xu, Lai Jiang, Chen Li, Zulin Wang, and Xiaoming Tao. Viewport-based CNN: A multi-task approach for assessing 360 video quality. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(4):2198–2215, 2022.
- [92] Yuecong Xu, Haozhi Cao, Kezhi Mao, Zhenghua Chen, Lihua Xie, and Jianfei Yang. Aligning correlation information for domain adaptation in action recognition. *IEEE Transactions on Neural Networks and Learning Systems*, 1(1):1–12, 2022.
- [93] Wufeng Xue, Xuanqin Mou, Lei Zhang, Alan C Bovik, and Xiangchu Feng. Blind image quality assessment using joint statistics of gradient magnitude and Laplacian features. *IEEE Transactions on Image Processing*, 23(11):4850–4862, 2014.
- [94] Bo Yan, Bahetiyaer Bare, and Weimin Tan. Naturalness-aware deep no-reference image quality assessment. *IEEE Transactions on Multimedia*, 21(10):2603–2615, 2019.
- [95] Qi Yang, Zhan Ma, Yiling Xu, Zhu Li, and Jun Sun. Inferring point cloud quality via graph similarity. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(6):3015–3029, 2022.
- [96] Wen Yang, Jinjian Wu, Shiwei Tian, Leida Li, Weisheng Dong, and Guangming Shi. Fine-Grained Image Quality Caption With Hierarchical Semantics Degradation. *IEEE Transactions on Image Processing*, 31(1):3578–3590, 2022.
- [97] Junjie Ye, Changhong Fu, Guangze Zheng, Ziang Cao, and Bowen Li. DarkLighter: Light up the darkness for UAV tracking. In *IEEE International Conference on Intelligent Robots and Systems (IROS)*, pages 3079–3085, 2021.
- [98] Zhenqiang Ying, Deepthi Ghadiyaram, and Alan Bovik. Telepresence video quality assessment. In *Springer European Conference on Computer Vision (ECCV)*, pages 327–347, 2022.
- [99] Zhenqiang Ying, Haoran Niu, Praful Gupta, Dhruv Mahajan, Deepthi Ghadiyaram, and Alan C. Bovik. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3575–3585, 2020.
- [100] Guangtao Zhai and Xiongkuo Min. Perceptual image quality assessment: A survey. *Springer Science China Information Sciences*, 63(11):1–52, 2020.
- [101] Guangtao Zhai, Xiaolin Wu, Xiaokang Yang, Weisi Lin, and Wenjun Zhang. A psychovisual quality metric in free-energy principle. *IEEE Transactions on Image Processing*, 21(1):41–52, 2012.
- [102] Lin Zhang, Lei Zhang, and Alan C Bovik. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing*, 24(8):2579–2591, 2015.
- [103] Weixia Zhang, Dingquan Li, Chao Ma, Guangtao Zhai, Xiaokang Yang, and Kede Ma. Continual learning for blind image quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1(1):1–15, 2022.
- [104] Weixia Zhang, Kede Ma, Jia Yan, Dexiang Deng, and Zhou Wang. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1):36–47, 2020.
- [105] Weixia Zhang, Kede Ma, Guangtao Zhai, and Xiaokang Yang. Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, 30(1):3474–3486, 2021.
- [106] Yonghua Zhang, Xiaoje Guo, Jiayi Ma, Wei Liu, and Jiawan Zhang. Beyond brightening low-light images. *Springer International Journal of Computer Vision*, 129(4):1013–1037, 2021.
- [107] Linru Zheng, Liquan Shen, Jianan Chen, Ping An, and Jun Luo. No-reference quality assessment for screen content images based on hybrid region features fusion. *IEEE Transactions on Multimedia*, 21(8):2057–2070, 2019.
- [108] Yannan Zheng, Weiling Chen, Rongfu Lin, Tiesong Zhao, and Patrick Le Callet. Uif: An objective quality assessment for underwater image enhancement. *IEEE Transactions on Image Processing*, 1:5456–5468, 2022.
- [109] Fei Zhou, Rongguo Yao, Bozhi Liu, and Guoping Qiu. Visual quality assessment for super-resolved images: Database and method. *IEEE Transactions on Image Processing*, 28(7):3528–3541, 2019.
- [110] Wei Zhou, Jiahua Xu, Qiuping Jiang, and Zhibo Chen. No-reference quality assessment for 360-degree images by analysis of multifrequency information and local-global naturalness. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(4):1778–1791, 2022.
- [111] Hancheng Zhu, Leida Li, Jinjian Wu, Weisheng Dong, and Guangming Shi. MetaIQA: Deep meta-learning for no-reference image quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14143–14152, 2020.