No-Reference Image Quality Assessment: Obtain MOS From Image Quality Score Distribution

Yixuan Gao, Xiongkuo Min, *Member, IEEE*, Yuqin Cao, Xiaohong Liu, *Member, IEEE*, and Guangtao Zhai, *Senior Member, IEEE*

***Abstract*— Recent image quality assessment (IQA) methods typically focus on predicting the mean opinion score (MOS) of image quality, ignoring the image quality score distribution. This distribution provides valuable information beyond the MOS, including the standard deviation of opinion scores (SOS) and opinion scores at different quality levels. This paper introduces a novel no-reference IQA method that predicts the image quality score distribution to estimate the MOS. The proposed method consists of three modules: a visual feature extraction module, a graph convolutional module, and a MOS prediction module. In the visual feature extraction module, a convolutional neural network is designed to extract both first- and second-order visual features of images. The graph convolutional module employs a graph convolutional network (GCN)-based mapper to map these visual features to the image quality score distribution by exploring correlations between quality labels. The MOS is then derived from the predicted image quality score distribution in the MOS prediction module. We are the first to jointly train the method using both the MOS and the image quality score distribution, enabling it to learn richer subjective information and improve prediction performance. To address the lack of the ground-truth image quality score distribution in some IQA databases, we propose to use a SOS assumption to generate a Gaussian-based image quality score distribution that better reflects subjective perception. Additionally, we design appropriate loss functions for training. Experimental results demonstrate that our method effectively predicts both the image quality score distribution and the MOS, outperforming most state-of-the-art IQA methods.**

***Index Terms*— Image quality assessment, image quality score distribution, GCN-based mapper, SOS assumption, loss functions.**

Received 1 August 2024; revised 23 September 2024; accepted 10 October 2024. Date of publication 24 October 2024; date of current version 13 February 2025. This work was supported in part by the National Natural Science Foundation of China under Grant 62271312, Grant 62132006, Grant 62225112, and Grant 62301310; in part by the National Key Research and Development Program of China under Grant 2021YFE0206700; in part by the Fundamental Research Funds for Central Universities, Shanghai Municipal Science and Technology Major Project under Grant 2021SHZDZX0102; in part by the Science and Technology Commission of Shanghai Municipality under Grant 22DZ2229005; and in part by Sichuan Science and Technology Program under Grant 2024NSFSC1426. This article was recommended by Associate Editor T. Chen. *(Corresponding authors: Xiongkuo Min; Guangtao Zhai.)*

Yixuan Gao, Xiongkuo Min, Yuqin Cao, and Guangtao Zhai are with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: gaoyixuan@sjtu.edu.cn; minxiongkuo@sjtu.edu.cn; caoyuqin@sjtu.edu.cn; zhaiguangtao@sjtu.edu.cn).

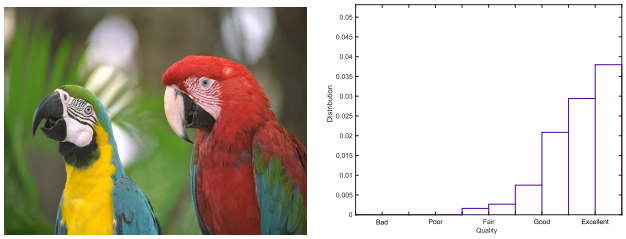
Xiaohong Liu is with the John Hopcroft Center (JHC) for Computer Science, Shanghai Jiao Tong University, Shanghai 200240, China (e-mail: xiaohongliu@sjtu.edu.cn).

Digital Object Identifier 10.1109/TCSVT.2024.3485684

I. Introduction

ONE of the most significant areas of image processing research is image quality assessment (IQA) [1], [2], [3], [4]. Generally, IQA can be divided into subjective IQA and objective IQA. Subjective IQA collects opinion scores on image quality from a large number of subjects using different quality labels. From these opinion scores, the mean opinion score (MOS) is derived, providing a general measure of perceptual image quality [5], [6], [7], [8], [9]. Furthermore, some subjective IQA methods use the image quality score distribution, known as the distribution of opinion scores, to offer a more comprehensive description of image quality [10], [11]. Objective IQA quantitatively evaluates image quality using mathematical models and algorithms. Depending on the amount of reference image information used, objective IQA methods are classified into fully-reference (FR) [12], [13], reduced-reference (RR) [14], [15], [16], and no-reference (NR) [17], [18], [19], [20], [21], [22] methods. NR IQA is particularly important due to the practical challenge of obtaining reference images, making it a focal point of contemporary research.

Conventional NR IQA methods extract handcrafted low-level image features and utilize regression models to predict image quality [23], [24], [25]. Recently, due to the strong learning capacity of the convolutional neural network (CNN), researchers have developed numerous CNN-based methods that predict image quality by extracting high-level visual features from images [26], [27], [28], [29], [30], [31]. Although notable advances have been made, most NR IQA methods focus primarily on predicting the MOS, neglecting the valuable information contained in the image quality score distribution. This distribution provides not only the MOS but also additional subjective information, such as the standard deviation of opinion scores (SOS) and opinion scores at different quality levels [10], [32], [33], [34]. The SOS reflects the consistency or disagreement in quality perception, with a smaller SOS suggesting more consistent and reliable quality scores and a larger SOS indicating significant differences in perception. Opinion scores at different quality levels can provide a more comprehensive description of image quality. Recently, several methods have been developed to predict the image quality score distribution [10], [11], [35], [36], [37]. For example, Talebi et al. introduced NIMA to predict the image quality score distribution [35]. Hosu et al. constructed



(a) 'Bird'. (b) The image quality score distribution of 'Bird'

Fig. 1. An image and its image quality score distribution.

the largest authentically distorted IQA database and proposed a method to predict the image quality score distribution [11]. Gao et al. used the alpha stable model to parameterize the image quality score distribution [10], [38]. In this paper, we propose a NR IQA method that extracts rich visual features from the image to predict the image quality score distribution, which is then used to calculate the MOS.

The international telecommunication union (ITU) recommends that subjects assess image quality using a five-level scale [39]. This scale comprises five quality labels: 'Bad', 'Poor', 'Fair', 'Good', and 'Excellent', representing a range of image quality from low to high. These quality labels are widely recognized and commonly used in subjective IQA [10], [11], [40], [41] and have influenced the prediction performance of objective IQA methods. We suggest that the image quality score distribution is closely related to these quality labels. For example, as shown in Fig. 1, the probability of 'Excellent' and 'Good' co-occurring to describe image quality is high, whereas the probability of 'Fair' and 'Excellent' co-occurring is very low. This indicates that quality labels are correlated rather than independently used to describe image quality. Moreover, we observe that different pairs of quality labels do not have the same probability of co-occurring to describe image quality, which reflects the image quality score distribution. Based on this analysis, we propose a graph convolutional module to map image features to the image quality score distribution by capturing correlations between quality labels. This module constructs a graph convolutional network (GCN) on quality labels using features extracted by the GloVe model [42] and designs an effective correlation matrix to guide information propagation between nodes, enabling the GCN to learn the correlations between quality labels. In this way, we construct an interdependent GCN-based mapper [43]. By applying the GCN-based mapper to image features, the proposed method can predict the image quality score distribution and subsequently calculate the MOS.

To enhance the ability of our method to learn richer subjective information and improve prediction performance, we jointly train the proposed method using both the MOS and the image quality score distribution. A significant challenge is the lack of the ground-truth image quality score distribution in some IQA databases. To address this issue and improve the applicability of our method, we develop various methods to generate the image quality score distribution and design suitable loss functions for training. For IQA databases that provide the MOS and SOS of image quality, such as the CSIQ [5], LIVE MD [40], LIVE Challenge [41], and CID2013 [44] databases, we directly generate the Gaussian-based image quality score distribution using the provided MOS and SOS. For IQA databases that provide only the MOS of image quality, such as the SPAQ [7] and VCLFER [45] databases, we first calculate a specific SOS for each image based on the SOS assumption [46] that accounts for consistency or disagreement in quality perception. The calculated SOS aligns better with subjective perception than directly assigning a uniform SOS to all images [47]. We then use the calculated SOS and the provided MOS to indirectly obtain the Gaussian-based image quality score distribution.

In conclusion, this paper proposes a novel method for predicting the image quality score distribution and subsequently calculating the MOS. The proposed method comprises three modules: a visual feature extraction module, a graph convolutional module, and a MOS prediction module. The visual feature extraction module uses a CNN with a backbone network, a channel attention module (CAM), and global pooling to extract image features. In the graph convolutional module, a GCN-based mapper is designed to learn the mapping between image features and the image quality score distribution. Finally, the MOS prediction module outputs both the MOS and the image quality score distribution.

Here is a summary of the contributions made in this paper.

- We propose a novel NR IQA method that predicts both the MOS and the image quality score distribution. This method innovatively uses a GCN to learn the mapping between image features and the image quality score distribution by capturing the correlations between quality labels.
- We propose training the method using both the MOS and the image quality score distribution, allowing it to learn richer subjective information and improve performance. To address the issue of some IQA databases lacking the ground-truth image quality score distribution, we introduce a method for generating the Gaussian-based image quality score distribution that better aligns with subjective perception.
- We conduct extensive experiments that demonstrate the superiority of the proposed method over state-of-the-art IQA methods in predicting both the MOS of image quality and the image quality score distribution.

The remaining sections of this paper are arranged as follows: In Section II, we introduce the proposed method in detail. In Section III, we verify the effectiveness and feasibility of the proposed method through extensive experiments. Finally, in Section IV, we conclude the paper.

II. Proposed Method

This section introduces the proposed NR IQA method, which includes a visual feature extraction module, a graph convolutional module, and a MOS prediction module. The framework of the method is shown in Fig. 2. In addition, we propose training the method using both the MOS and the image quality score distribution, with specifically designed loss functions to enhance performance.

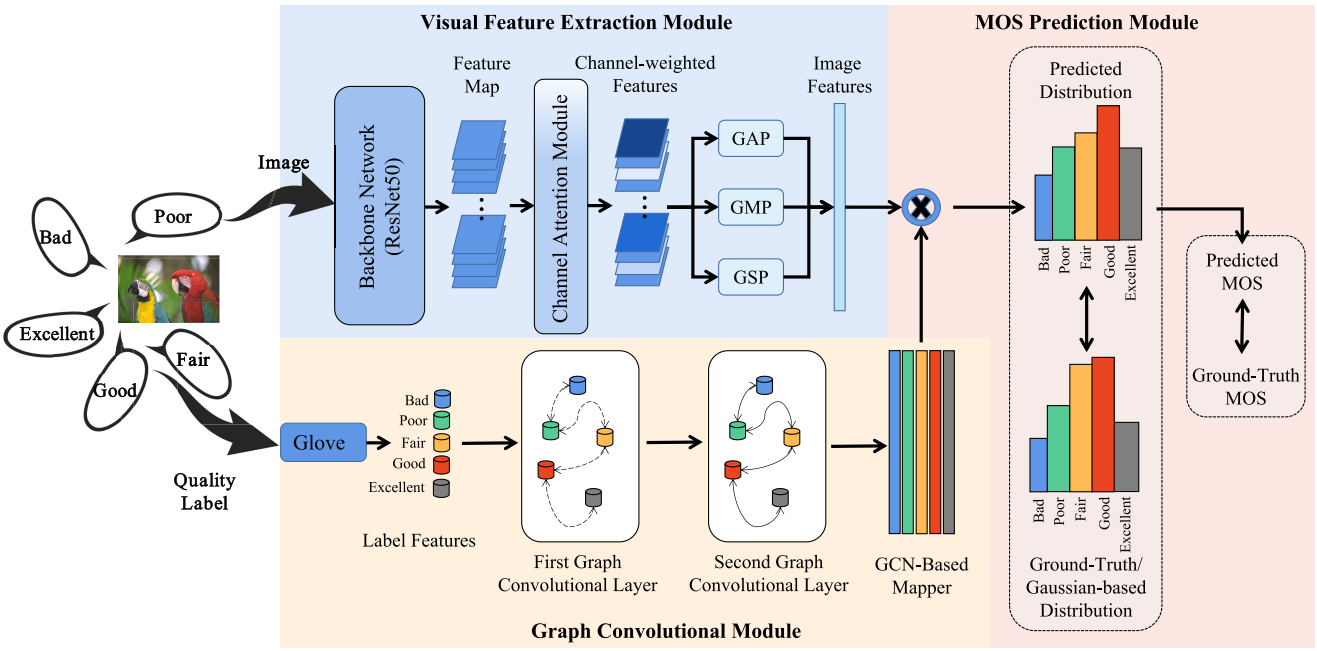


Fig. 2. Framework of the proposed method. The blue background represents the visual feature extraction module, the yellow background indicates the graph convolutional module, and the pink background illustrates the MOS prediction module and the training process (double-directional arrows). GAP means global average pooling. GMP means global maximum pooling. GSP means global second-order pooling. Dotted lines in the first graph convolutional layer and solid lines in the second graph convolutional layer indicate the correlations between quality labels. The blue circle with a black multiplication symbol indicates that the extracted image features are processed by the GCN-based mapper.

A. Visual Feature Extraction Module

Due to subjective differences, individuals tend to focus on different visual features of the same image. To better align with human perception of image quality, representative image features are extracted in the visual feature extraction module. This module employs ResNet50 [48], a backbone network renowned for its robust learning capabilities and superior performance in visual tasks, to extract image features:

$$\mathbf{F} = ResNet(\mathbf{I}), \tag{1}$$

where $\mathbf{I}$ is the input image and $\mathbf{F}$ is the extracted image features. It is important to note that ResNet50 used in this paper excludes the final global average pooling (GAP) and fully-connected (FC) layer. Therefore, $\mathbf{F}$ consists of the image features extracted from the last convolutional layer of ResNet-50. To emphasize significant image features, CAM [49] is employed to assign greater weights to these features:

$$\mathbf{F}' = CAM(\mathbf{F}), \tag{2}$$

where $\mathbf{F}'$ represents the channel-weighted features.

Subsequently, first- and second-order visual features are extracted. First-order visual features are obtained using GAP and global maximum pooling (GMP):

$$\mathbf{F}'_{GAP} = GAP(\mathbf{F}'), \ \mathbf{F}'_{GMP} = GMP(\mathbf{F}'). \tag{3}$$

Global pooling provides a comprehensive representation of image features, which helps prevent overfitting and enhances the consistency of feature maps [50]. Second-order visual features, which have been shown to be effective in visual recognition tasks [51], are extracted using global second-order pooling (GSP) [52]:

$$\mathbf{F}'_{GSP} = GSP(\mathbf{F}'). \tag{4}$$

This multi-order feature extraction captures intricate relationships between features, providing a more comprehensive representation of the image [53].

Finally, by concatenating $\mathbf{F}'_{GAP}$, $\mathbf{F}'_{GMP}$, and $\mathbf{F}'_{GSP}$, the comprehensive visual features of the image $\mathbf{I}$ are obtained:

$$\mathbf{F}'' = \mathbf{F}'_{GAP} \oplus \mathbf{F}'_{GMP} \oplus \mathbf{F}'_{GSP}, \tag{5}$$

where $\mathbf{F}''$ is the image features extracted by the visual feature extraction module, and $\oplus$ denotes the concatenation operation along the feature channels.

B. Graph Convolutional Module

We design a graph convolutional module to map the extracted image features to the image quality score distribution by capturing the correlations between quality labels.

1) Label Feature Extraction: In subjective IQA, five quality labels are commonly used to describe image quality: 'Bad', 'Poor', 'Fair', 'Good', and 'Excellent'. These quality labels guide subjects in assessing image quality. Although these labels represent different levels of image quality, they are not independent and often co-occurr to describe image quality. To effectively capture and explore the correlations between these quality labels, we first extract label features using the GloVe model [42], which is a count-based statistical word representation model. The extracted label features are

as follows:

$$\begin{aligned} \mathcal{L}_B &= GloVe(\text{bad}),\ \ \mathcal{L}_{\text{P}} = GloVe(\text{poor}),\\ \mathcal{L}_F &= GloVe(\text{fair}),\ \ \mathcal{L}_{\text{G}} = GloVe(\text{good}),\\ \mathcal{L}_E &= GloVe(\text{excellent}). \end{aligned} \tag{6}$$

2) Correlation Matrix: The GCN needs to propagate information between nodes (labels) through a correlation matrix. We construct an initial correlation matrix for the five quality labels in a data-driven manner. Specifically, in an IQA database, we first calculate the number of times two labels simultaneously describe the same image: M_{ij}, where $i, j \in \{1, 2, 3, 4, 5\} = \{\text{bad, poor, fair, good, excellent}\}$. This computation results in the label co-occurrence matrix:

$$\boldsymbol{M} = \begin{bmatrix} M_{11} & M_{12} & \cdots & M_{15} \\ M_{21} & M_{22} & \cdots & M_{25} \\ \vdots & \vdots & \ddots & \vdots \\ M_{51} & M_{52} & \cdots & M_{55} \end{bmatrix}. \tag{7}$$

Then, the conditional probability matrix can be written as:

$$\boldsymbol{P} = \begin{bmatrix} P_{11} & P_{12} & \cdots & P_{15} \\ P_{21} & P_{22} & \cdots & P_{25} \\ \vdots & \vdots & \ddots & \vdots \\ P_{51} & P_{52} & \cdots & P_{55} \end{bmatrix}, \tag{8}$$

where $P_{ij} = P(j|i) = M_{ij}/M_i$ is the probability of label j occurring when label i appears, in which M_i is the number of times label i occurs. A smaller value of P_{ij} indicates a weaker correlation between the two quality labels, while a larger value indicates a stronger correlation. Then, the matrix $\boldsymbol{P}$ is binarized [43], [54]:

$$A_{ij} = \begin{cases} 0, & P_{ij} < \tau \\ 1, & P_{ij} \geq \tau, \end{cases} \tag{9}$$

where τ is the threshold. Here, $\boldsymbol{A} = [A_{ij}] \in \mathbb{R}^{5\times 5}$ is the binary correlation matrix. Finally, we re-weight $\boldsymbol{A}$ and obtain the final correlation matrix $\widetilde{\boldsymbol{A}} = [\widetilde{A}_{ij}] \in \mathbb{R}^{5\times 5}$ [43], [54]:

$$\widetilde{A}_{ij} = \begin{cases} A_{ij} \times (p/\sum_{j=1, j\neq i}^{5} A_{ij}), & i \neq j \\ 1 - p, & i = j, \end{cases} \tag{10}$$

where p is the weight.

3) GCN-Based Mapper: Using the constructed correlation matrix, the GCN can learn the correlations among nodes (labels) by propagating information between them. The GCN is applied to map the label features (*i.e.* Eq. (6)) into an interdependent mapper, which is then used on the extracted image features. Specifically, the GCN consists of two graph convolutional layers [55] and one LeakyReLU layer [56]. The first graph convolutional layer takes the label features $\mathcal{L} = [\mathcal{L}_B; \mathcal{L}_P; \mathcal{L}_F; \mathcal{L}_G; \mathcal{L}_E] \in \mathbb{R}^{N\times D}$ and the corresponding correlation matrix $\widetilde{\boldsymbol{A}} \in \mathbb{R}^{N\times N}$ as inputs and learns new label features $\mathcal{L}' \in \mathbb{R}^{N\times D'}$ using a layer-wise propagation rule $f(\cdot, \cdot)$:

$$\mathcal{L}' = f(\mathcal{L}, \widetilde{\boldsymbol{A}}), \tag{11}$$

where $N = 5$ represents the number of labels, D represents the dimension of the label features, and D' represents the dimension of the new label features. The output $\mathcal{L}'$ is then input into a LeakyReLU layer:

$$\mathcal{L}' = LeakyReLU(\mathcal{L}'). \tag{12}$$

Finally, the second graph convolutional layer takes the label features $\mathcal{L}'$ and the correlation matrix $\widetilde{\boldsymbol{A}}$ as inputs, generating new label features $\mathcal{L}'' \in \mathbb{R}^{N\times D''}$ using the propagation rule $f(\cdot, \cdot)$:

$$\mathcal{L}'' = f(\mathcal{L}', \widetilde{\boldsymbol{A}}), \tag{13}$$

where D'' represents the dimension of the new label features, which is equal to the dimension of $\mathbf{F}''$. $\mathcal{L}''$ is the proposed GCN-based mapper.

C. MOS Prediction Module

By applying the GCN-based mapper $\mathcal{L}''$ to the extracted image features $\mathbf{F}''$, we can predict the image quality score distribution:

$$\hat{\mathbf{P}} = \{\hat{p}_1, \hat{p}_2, \cdots, \hat{p}_N\} = \mathcal{L}''\mathbf{F}''. \tag{14}$$

The MOS can be calculated as follows:

$$\widehat{\text{MOS}} = \sum_{i=1}^{N} l_i \hat{p}_i, \tag{15}$$

where l_i denotes the quality score assigned to the i-th quality label.

D. Training and Loss Function

To enable our method to learn richer subjective information and improve prediction performance, we train the proposed method using both the MOS and the image quality score distribution. The loss function designed in this paper consists of a distribution-based loss function and a MOS-based loss function.

1) Distribution-Based Loss Function: When designing the distribution-based loss function, we consider three categories of IQA databases. For IQA databases providing the image quality score distribution, the distribution-based loss function can be expressed as follows [57]:

$$\text{Loss}_1 = \sqrt{\frac{1}{N}\sum_{k=1}^{N} |\sum_{i=1}^{k} p_i - \sum_{i=1}^{k} \hat{p}_i|^2}, \tag{16}$$

where $\mathbf{P} = \{p_1, p_2, \cdots, p_N\}$ is the ground-truth image quality score distribution.

Previous works [35], [36] assumed that the ground-truth image quality score distribution follows a Gaussian distribution. In this paper, we obtain the Gaussian-based image quality score distribution for IQA databases that provide both the MOS and the SOS:

$$p_i = \frac{1}{\text{SOS}\sqrt{2\pi}} e^{\frac{-(s_i - \text{MOS})^2}{2\text{SOS}^2}},\ \ i = 1, 2, \ldots, N. \tag{17}$$

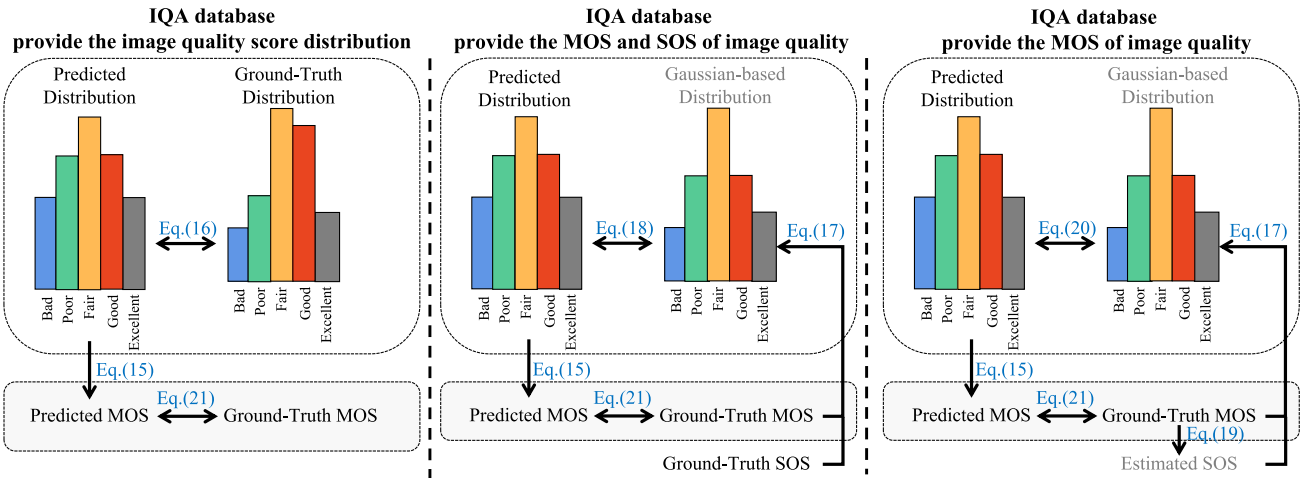


Fig. 3. Training methods. From left to right, the training methods are tailored for three types of IQA databases: those that provide the image quality score distribution, those that provide the MOS and SOS, and those that provide only the MOS. The white background represents the process of training the method using the distribution-based loss function. The gray background represents the process of training the method using the MOS-based loss function. Equations are highlighted in blue text. Gray text indicates information that is not directly provided by the database but generated using the corresponding equations.

By substituting Eq. (17) into Eq. (16), we use the following distribution-based loss function to train the method:

$$\text{Loss}_1 = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left|\sum_{i=1}^{k}\left(\frac{e^{\frac{-(s_i-\text{MOS})^2}{2\text{SOS}^2}}}{\text{SOS}\sqrt{2\pi}}\right)-\sum_{i=1}^{k}\hat{p}_i\right|^2}. \tag{18}$$

For IQA databases that provide only the MOS of image quality, we use a SOS assumption to generate the Gaussian-based image quality score distribution. Specifically, we first calculate the SOS based on the SOS assumption [46]:

$$\text{SOS}^2 = a(-\text{MOS}^2 + (s_1 + s_N)\text{MOS} - s_1 s_N), \tag{19}$$

where a is an empirical parameter, MOS is the ground-truth MOS, and $[s_1, s_N]$ represents the quality scale. Zeng et al. [47] simulated the Gaussian-based image quality score distribution by assigning the same SOS to all images. However, high-quality and low-quality images tend to have more concentrated subjective scores with smaller SOS values, while medium-quality images have more dispersed subjective scores with larger SOS values [46]. This may be because high-quality images are usually clear and detailed, and low-quality images are visibly damaged, making them easy to rate consistently. However, medium-quality images have less distinct quality features, resulting in larger SOS values. Eq. (19) reveals that the SOS is smaller when the MOS is high or low and larger when the MOS is in the middle range. Therefore, the Gaussian-based image quality score distribution generated from Eq. (19) is more aligned with subjective perception.

Then, the calculated SOS and the provided MOS are substituted into Eqs. (17) and (16), the following distribution-based loss function is used to train the method:

$$\text{Loss}_1 = \left(\frac{1}{N}\sum_{k=1}^{N}\left|\sum_{i=1}^{k}\left(\frac{e^{\frac{-(s_i-\text{MOS})^2}{2\left(a(-\text{MOS}^2+(s_1+s_N)\text{MOS}-s_1s_N)\right)^2}}}{\sqrt{2\pi a(-\text{MOS}^2+(s_1+s_N)\text{MOS}-s_1s_N)}}\right)-\sum_{i=1}^{k}\hat{p}_i\right|^2\right)^{\frac{1}{2}}. \tag{20}$$

2) MOS-Based Loss Function: All IQA databases provide the ground-truth MOS of image quality. Therefore, the MOS-based loss function used in this paper is the norm-in-norm loss with mean normalization [58]:

$$\text{Loss}_2 = \left(\frac{\text{MOS} - \text{MOS}_{mean}}{\text{MOS}_{max} - \text{MOS}_{min}} - \frac{\widehat{\text{MOS}} - \widehat{\text{MOS}}_{mean}}{\widehat{\text{MOS}}_{max} - \widehat{\text{MOS}}_{min}}\right)^2, \tag{21}$$

where MOS_{mean}, MOS_{max}, and MOS_{min} are the mean, maximum, and minimum of ground-truth MOSs in a training batch. Similarly, $\widehat{\text{MOS}}_{mean}$, $\widehat{\text{MOS}}_{max}$, and $\widehat{\text{MOS}}_{min}$ are the mean, maximum, and minimum of the predicted MOSs in a training batch.

In this paper, the distribution-based loss function Loss_1 and the MOS-based loss function Loss_2 are combined to train the method:

$$\text{Loss} = \frac{1}{M}\sum_{m=1}^{M}\text{Loss}_1^m + \alpha\frac{1}{M}\sum_{m=1}^{M}\text{Loss}_2^m, \tag{22}$$

where M is the training batch, Loss_1^m and Loss_2^m represent the Loss_1 and Loss_2 of the m-th image in the training batch. The parameter α is a constant greater than zero. Fig. 3 provides a summary of the training methods developed for different IQA databases. From left to right, the figure illustrates the training methods designed for IQA databases that provide the image quality score distribution, those that provide the MOS and SOS of image quality, and those providing only the MOS of image quality.

III. Experiments

In the following section, we conduct experiments to verify the feasibility and effectiveness of the proposed method.

A. Database

We select eight IQA databases to validate our method, including the SJTU IQSD [10], KonIQ-10K (1024×768 pixel)

TABLE I
INFORMATION ABOUT IQA DATABASES. '# DISTORTED' IS THE NUMBER OF DISTORTED IMAGES. '# REFERENCE' DENOTES THE NUMBER OF REFERENCE IMAGES

Category	Database	# Distorted	# Reference	Scenario	Annotation	Quality Scale	a
First	SJTU IQSD	779	29	Synthetic	Distribution	[0,100]	0.1449
	KonIQ-10K	10073	-	Authentic	Distribution	1,2,3,4,5	0.0907
Second	CSIQ	866	30	Synthetic	MOS and SOS	[0,1]	0.0656
	LIVE MD	450	15	Synthetic	MOS and SOS	[0,100]	0.1322
	LIVE Challenge	1162	-	Authentic	MOS and SOS	[0,100]	0.1841
	CID2013	474	-	Authentic	MOS and SOS	[0,100]	0.1649
Third	VCLFER	552	23	Synthetic	MOS	[0,100]	-
	SPAQ	11125	-	Authentic	MOS	[0,100]	-

[11], CSIQ [5], LIVE MD [40], LIVE Challenge [41], CID2013 [44], SPAQ [7], and VCLFER [45] databases. These databases are classified into three categories, and experiments are conducted on each category to evaluate the performance of our method. Information about these databases is summarized in Table I.

1) Distribution: The first category of IQA databases provides complete subjective opinion scores for image quality, known as the image quality score distribution. For example, the SJTU IQSD [10] and KonIQ-10K [11] databases.

The SJTU IQSD database is developed from the renowned LIVE database. It has 779 distorted images and 29 reference images, with distortion types including JPEG2000 (JP2K), JPEG, white noise (WN), Gaussian blur (Gblur), and fast-fading Rayleigh (FF). The authors invited 206 subjects to rate these 808 images on a quality scale of [0, 100]. After screening, each image has 187 valid subjective opinion scores. These scores are publicly available as the image quality score distribution for each image.

The KonIQ-10K database is one of the largest authentically distorted IQA databases comprising 10,073 images. This database collected 1.2 million reliable quality scores through crowd-sourcing. The quality score is divided into five quality levels, corresponding to scores of 1, 2, 3, 4, and 5, which represent image quality from bad to excellent. In addition, the authors have released the number of scores at each quality level for each image, which constitutes the image quality score distribution.

2) MOS and SOS: The second category of IQA databases provides the MOS and SOS of image quality, such as the CSIQ [5], LIVE MD [40], LIVE Challenge [41], and CID2013 [44] databases.

The CSIQ database has 866 distorted images and 30 reference images. The distortion types include JPEG, JPEG2K, overall contrast reduction, additive Gaussian noise, and Gblur. The differential MOS (DMOS) of this database was obtained from approximately 5,000 opinion scores provided by 25 subjects, with a value range of [0,1]. The SOS of image quality is also available.

The LIVE MD database is a multiply distorted IQA database. The database has a total of 450 distorted images and 15 reference images, involving two types of multiple distortions. The first type of multiple distortion is image storage, where the image is first blurred and then compressed by a JPEG encoder. The second type of multiple distortion is the camera image acquisition process, where the image is first blurred due to narrow depth of field or other defocus and then corrupted by white Gaussian noise to simulate sensor noise. The database provides both the MOS and SOS of image quality.

The LIVE Challenge database has 1,162 authentically distorted images captured from many different portable electronics. The authors utilized Amazon's crowdsourcing system to collect subjective opinion scores. Each image was viewed and rated online by an average of 175 subjects on a continuous quality scale of [0,100]. The MOS and SOS of image quality from this subjective assessment are available.

The CID2013 database has 474 real photographic images captured by different digital cameras. The authors invited a total of 188 subjects to rate these images. Each image was viewed and rated by an average of 30 subjects on a continuous quality scale of [0,100]. The authors have released the MOS and SOS of image quality.

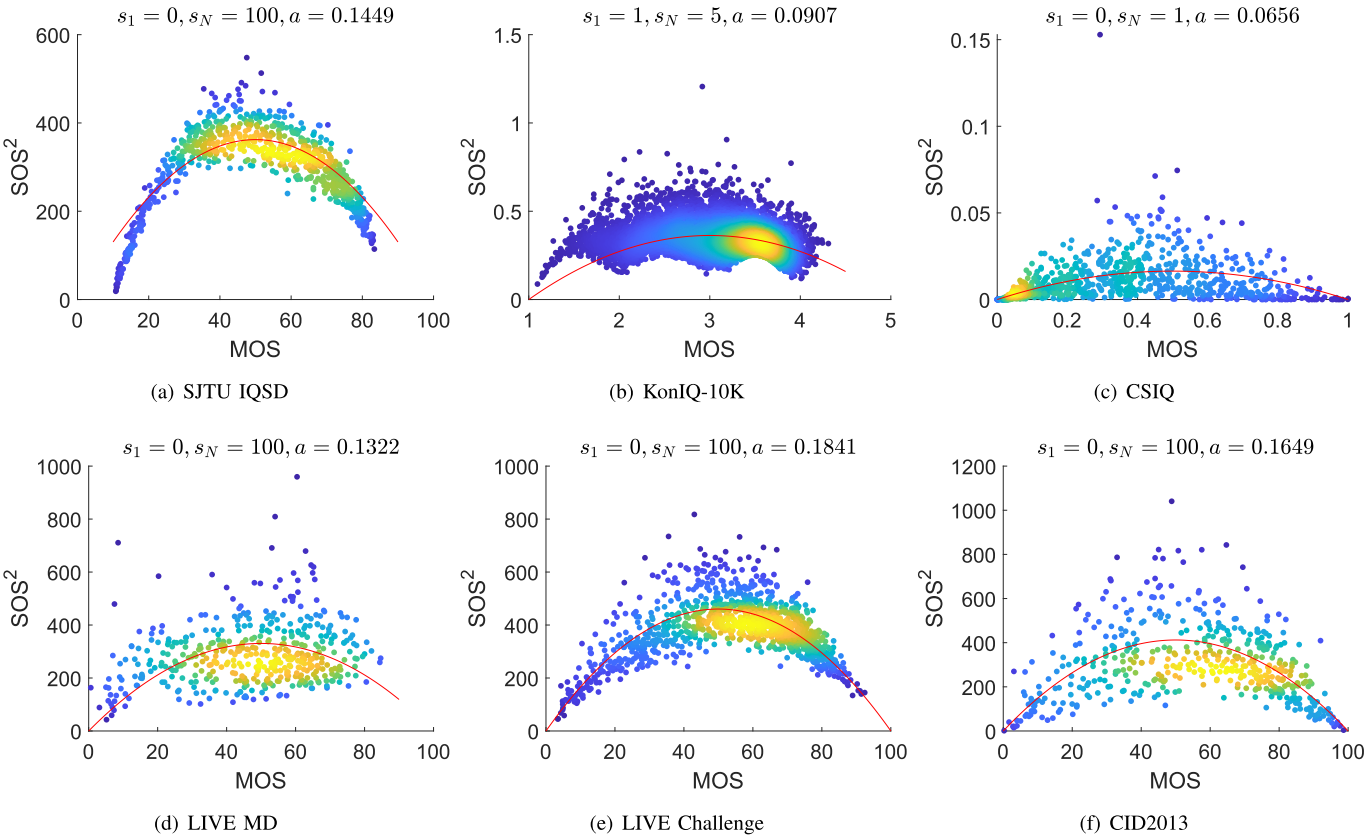
3) MOS: The third category of IQA databases provides only the MOS for image quality, such as the SPAQ [7] and VCLFER [45] databases.

The SPAQ database is currently the largest authentically distorted database, consisting of 11,125 images taken by 66 smartphones. The author invited over 600 subjects to rate the quality of these images on a continuous quality scale of [0, 100]. Only the MOS is provided.

The VCLFER database contains 575 images. Among these images, 23 are reference images. This database has four different types of distortions, including average white Gaussian noise (AWGN), Gblur, JPEG2K, and JPEG. Each distortion type has six quality levels. 118 subjects were required to rate the quality of these images on a continuous quality scale of [0,100]. The authors release only the MOS of image quality.

B. Experimental Setup

We first calculate the value of a for the first and second categories of IQA databases according to Eq. (19). The six figures in Fig. 4 show scatter plots and fitted curves of MOS and SOS^2 for the SJTU IQSD, KonIQ-10K, CSIQ, LIVE MD, LIVE Challenge, and CID2013 databases, respectively. As shown in Fig. 4, high-quality and low-quality images tend to have smaller SOS values, while medium-quality images have larger SOS values. From the fitted curve, the value of a for each database is computed and summarized in the



(a) SJTU IQSD (b) KonIQ-10K (c) CSIQ

(d) LIVE MD (e) LIVE Challenge (f) CID2013

Fig. 4. Scatter plots and fitted curves of MOS and SOS^2 for the SJTU IQSD, KonIQ-10K, CSIQ, LIVE MD, LIVE Challenge, and CID2013 databases. Scatter points represent images, with the color from dark to light indicating an increase in the number of images. The red curves are fitted curves of MOS and SOS^2 according to Eq. (19). s_1, s_N, and a for each database are shown above each figure.

TABLE II

PERFORMANCE COMPARISON OF THE PROPOSED METHOD AND COMPETING METHODS FOR PREDICTING THE MOS OF IMAGE QUALITY ON THE FIRST CATEGORY OF IQA DATABASES, *i.e.* SJTU IQSD AND KONIQ-10K DATABASES. THE BEST PERFORMANCES ARE IN BOLD

Method	SJTU IQSD			KonIQ-10K		
	SRCC	PLCC	RMSE	SRCC	PLCC	RMSE
BRISQUE	0.8821	0.8857	9.8369	0.6976	0.7010	0.3941
NIQE	0.9225	0.9265	8.7210	0.6746	0.6731	0.4097
BMPRI	0.9200	0.9285	8.3407	0.4269	0.4246	0.5110
Kang	0.8926	0.9049	9.5157	0.5711	0.5071	0.4993
DIQaM-NR	0.9250	0.9414	8.2896	0.7901	0.7734	0.4201
WaDIQaM-NR	0.9159	0.9251	9.1380	0.8484	0.8452	0.3272
SGDNet	0.9159	0.9183	8.2631	0.8840	0.9028	0.2417
DBCNN	0.9349	0.9439	7.8040	0.8594	0.8787	0.2743
HyperIQA	0.9424	0.9502	7.2852	0.9040	0.9163	0.2250
UNIQUE	0.9584	0.9603	5.4269	0.8901	0.8942	0.2452
DACNN	0.9470	0.9538	6.5960	0.8956	0.9121	0.2309
GraphIQA	0.9312	0.9425	7.9065	0.8389	0.8609	0.3548
NIMA	0.9316	0.9444	7.2501	0.7803	0.7868	0.3428
Gao	0.9451	0.9567	6.3271	0.9045	0.9185	0.2185
StairIQA	0.9336	0.9518	6.2685	0.9053	0.9167	0.2070
TSNIQA	0.9077	0.9287	9.4351	0.8654	0.8337	0.3279
TOPIQ	0.9564	0.9621	5.3781	0.9068	0.9136	0.2119
Proposed	**0.9622**	**0.9679**	**5.2706**	**0.9105**	**0.9247**	**0.2098**

last column of Table I. The mean value of a across the six IQA databases is 0.1304, which is set as the value of a for the third category of IQA databases. By substituting $a = 0.1304$ into Eq. (20), the Gaussian-based image quality score distribution can be used for training on the third category of IQA databases.

For the experiment, we randomly partition the database. The training set comprises 80% of the images, and the test set consists of the remaining 20%. For databases with synthetically distorted images, we partition them into training and test sets based on reference images. To minimize bias from the randomness of the partitioning, we repeat the above

TABLE III

PERFORMANCE COMPARISON OF THE PROPOSED METHOD AND COMPETING METHODS FOR PREDICTING THE MOS OF IMAGE QUALITY ON THE SECOND CATEGORY OF IQA DATABASES, *i.e.* CSIQ, LIVE MD, LIVE CHALLENGE, AND CID2013 DATABASES. THE BEST PERFORMANCES ARE IN BOLD

Method	CSIQ		LIVE MD		LIVE Challenge		CID2013	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE	0.7433	0.8014	0.8754	0.9009	0.5814	0.6039	0.5318	0.4791
NIQE	0.7722	0.8314	0.8969	0.9145	0.5803	0.5987	0.4813	0.4323
BMPRI	0.7343	0.7853	0.7837	0.8151	0.3724	0.3919	0.5293	0.5293
Kang	0.6561	0.7484	0.6455	0.7414	0.6612	0.6239	0.5254	0.5094
DIQaM-NR	0.8797	0.8868	0.9140	0.9319	0.5950	0.5936	0.7615	0.7528
WaDIQaM-NR	0.8222	0.7955	0.8982	0.8987	0.6544	0.6597	0.8357	0.8265
SGDNet	0.8250	0.8542	-	-	0.8115	0.8475	0.8280	0.8390
DBCNN	0.9166	0.9288	0.9061	0.9222	0.8314	0.8545	0.8710	0.8850
HyperIQA	0.9206	0.9111	**0.9500**	0.9116	0.8453	0.8589	0.8937	0.9027
UNIQUE	0.9193	0.9316	0.8013	0.8269	0.8345	0.8507	0.9081	0.9125
DACNN	0.9085	0.9409	0.8750	0.9286	0.8485	0.8602	0.8891	0.8949
GraphIQA	0.9200	0.9304	0.9470	0.9332	0.8079	0.8335	0.8824	0.8743
NIMA	0.9126	0.9283	0.9100	0.9272	0.7811	0.8124	0.7543	0.7734
Gao	0.9075	0.9200	0.9181	0.9054	0.7706	0.8012	0.8292	0.8448
StairIQA	0.8302	0.8642	0.9207	0.9150	0.8460	0.8685	0.8686	0.8710
TSNIQA	0.9011	0.9145	0.8972	0.9405	0.8117	0.8167	0.8275	0.8449
TOPIQ	0.8888	0.9005	0.9039	0.9208	0.8494	0.8637	0.9025	0.9071
Proposed	**0.9362**	**0.9429**	0.9478	**0.9431**	**0.8550**	**0.8731**	**0.9095**	**0.9211**

TABLE IV

PERFORMANCE COMPARISON OF THE PROPOSED METHOD AND SOME COMPETING METHODS FOR PREDICTING THE MOS OF IMAGE QUALITY ON THE THIRD CATEGORY OF IQA DATABASES, *i.e.* SPAQ AND VCLFER DATABASES. THE BEST PERFORMANCES ARE IN BOLD

Method	SPAQ		VCLFER	
	SRCC	PLCC	SRCC	PLCC
BRISQUE	0.7075	0.708	0.9092	0.9070
NIQE	0.7002	0.7010	0.8951	0.8904
BMPRI	0.6083	0.6110	0.8147	0.8215
DBCNN	0.8847	0.8892	0.8514	0.8101
HyperIQA	0.7528	0.7564	0.8571	0.7002
DACNN	0.8870	0.8912	0.9233	0.9156
GraphIQA	0.7690	0.7670	0.8229	0.7405
NIMA	0.7621	0.7666	0.9414	0.9410
Gao	0.9099	0.9131	0.9596	0.9542
StairIQA	0.9078	0.9113	0.9313	0.9324
TOPIQ	0.8991	0.9067	0.9481	0.9456
Proposed	**0.9121**	**0.9175**	**0.9613**	**0.9596**

partitioning procedure ten times and report the mean result. More specifically, we train our method using the Adam optimizer [59] with a learning rate of 10^{-4}. During training and testing, the image size is set to 448×448 pixels. The GloVe model used in this paper is 50-dimensional. The training batch size M is set to 8. The threshold τ is 0.05. The parameter p is set to 0.25, α is set to 2.5, and D' is set to 512. ResNet50 is pre-trained on the ImageNet database [60].

C. Competing Method

We compare the performance of the proposed method with some competing IQA methods for predicting the MOS of image quality. The methods include BRISQUE [23], NIQE [24], BMPRI [25], Kang et al. [61], (Wa)DIQaM-N [62], SGDNet [63], DBCNN [64], HyperIQA [65], UNIQUE [66], DACNN [27], GraphIQA [67], NIMA [35], Gao et al. [37], StairIQA [68], TSNIQA [69], and TOPIQ [70]. TOPIQ can be employed as either a FR or NR IQA method. To ensure a fair comparison, this paper evaluates the prediction performance of TOPIQ solely as a NR IQA method. We analyze the prediction performance using three criteria: Spearman rank correlation coefficient (SRCC), Pearson linear correlation coefficient (PLCC), and root mean squared error (RMSE). Better prediction performance is indicated by higher SRCC and PLCC (close to 1) and lower RMSE (close to 0). Additionally, we compare the performance of the proposed method with some competing methods for predicting the image quality score distribution, including NIMA [35], Liu et al. [36], IQSD-Alpha [10], KonCept512 [11], and Gao et al. [37]. We analyze the prediction performance using the following three criteria: Jensen-Shannon distance (JSD), earth mover's distance (EMD), RMSE, intersection, and cosine similarity. Higher intersection and cosine (closer to 1) indicate better prediction performance, while lower JSD, EMD, and RMSE (closer to 0) indicate better prediction performance.

D. Performance Comparison

We first compare the performance of the proposed method with competing methods for predicting the MOS, as reported in Tables II, III, and IV. Table II shows the performance results for predicting the MOS of image quality on the SJTU IQSD and KonIQ-10K databases. The results show that our proposed method achieves the best prediction performance, demonstrating its effectiveness in obtaining the MOS of image quality by predicting the image quality score distribution. Table III shows the performance for predicting the MOS of image quality on the CSIQ, LIVE MD, LIVE Challenge, and CID2013 databases, where our proposed method also achieves the best prediction performance on most databases. Table IV shows the performance results for predicting the MOS of image quality on the SPAQ and VCLFER databases, where our

TABLE V

PERFORMANCE COMPARISON OF THE PROPOSED METHOD AND COMPETING METHODS FOR PREDICTING THE IMAGE QUALITY SCORE DISTRIBUTION. THE BEST PERFORMANCES ARE IN BOLD

Method	SJTU IQSD					KonIQ-10K				
	JSD	EMD	RMSE	Intersection	Cosine	JSD	EMD	RMSE	Intersection	Cosine
NIMA	0.0278	0.0742	**0.0425**	0.8436	0.9352	0.0415	0.0835	0.1087	0.8073	0.9267
Liu	0.0423	0.1641	0.0697	0.7032	0.8498	0.0214	0.0952	0.1045	0.7950	0.9167
IQSD-Alpha	0.0518	0.1034	0.0653	0.7215	0.8762	-	-	-	-	-
KonCept512	0.0743	0.1435	0.0623	0.7345	0.8724	0.0624	0.1074	0.1214	0.7747	0.9032
Gao	0.0239	0.0687	0.0651	0.8529	0.9507	0.0213	0.0604	0.0719	0.8687	0.9622
Proposed	**0.0145**	**0.0601**	0.0561	**0.8873**	**0.9685**	**0.0201**	**0.0586**	**0.0693**	**0.8731**	**0.9645**

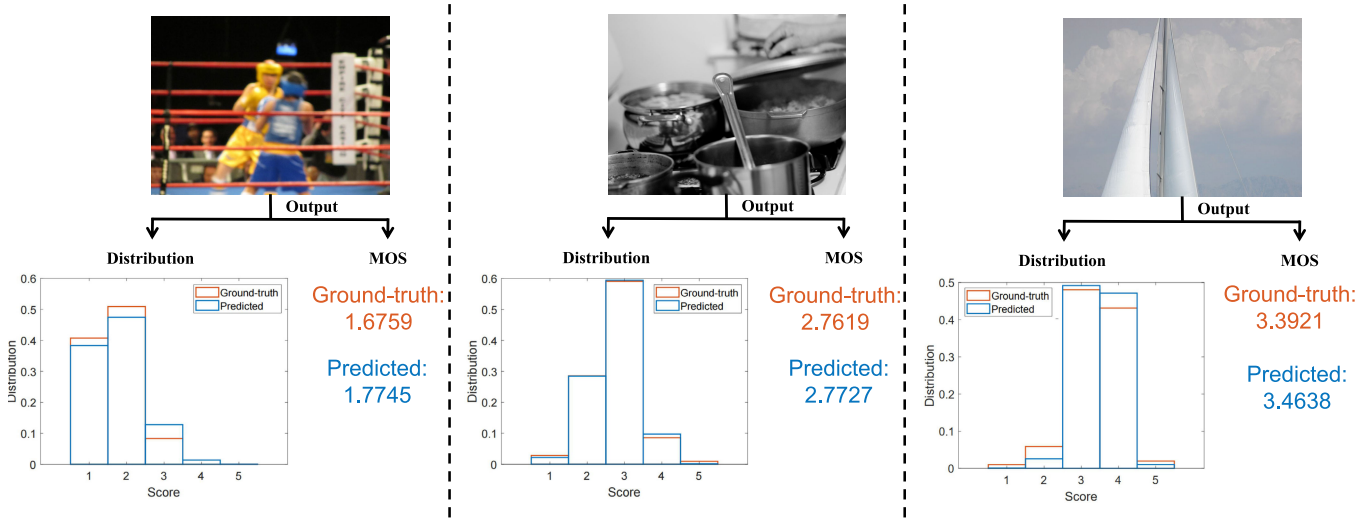


Fig. 5. Image quality score distribution and MOS predicted by our method. Blue represents the predicted results, and red represents the ground truth.

proposed method consistently outperforms the other methods. Tables III and IV indicate that it is effective to obtain the MOS of image quality by predicting the Gaussian-based image quality score distribution.

In addition, we compare the performance of the proposed method and competing methods in predicting the image quality score distribution, as shown in Table V. The results indicate that our method performs best in predicting the image quality score distribution on the KonIQ-10K database and remains highly competitive on the SJTU IQSD database. In summary, the proposed method demonstrates outstanding performance in predicting both the MOS of image quality and the image quality score distribution. Fig. 5 illustrates the predicted results of the proposed method for three images from the KonIQ-10K database, including the predicted image quality score distribution and MOS. The predicted results (blue) are highly consistent with the ground-truth image quality score distributions and MOSs (red).

E. Cross-Database Validation

Robustness is a crucial criterion for evaluating the effectiveness of a method. In this subsection, we conduct cross-database validations on the SJTU IQSD, KonIQ-10K, CSIQ, CID2013, SPAQ, and VCLFER databases to test the robustness of the proposed method and several competing methods. Specifically, the method is trained on one database and then tested on others to evaluate its performance in predicting the MOS. The results are shown in Table VI. When trained on the SJTU IQSD and CSIQ databases, the proposed method achieves the best performance in predicting the MOS on the KonIQ-10K database and shows competitive performance on the VCLFER database. When trained on the KonIQ-10K database, the proposed method achieves the best performance in predicting the MOS on the SJTU IQSD and SPAQ databases. When trained on the CID2013 database, the proposed method achieves the best performance in predicting the MOS on the SPAQ database and shows competitive performance on the SJTU IQSD database. In conclusion, the proposed method consistently outperforms most competing IQA methods across different databases, demonstrating robustness in cross-database validations. This confirms the effectiveness of our method.

F. Individual Distortions

The SJTU IQSD database is developed from the well-known LIVE database. It contains 29 reference images and 779 distorted images with distortion types including JP2K, JPEG, WN, Gblur, and FF. We evaluate the proposed method and several competing methods based on their prediction performance for individual distortions within the SJTU IQSD database. The results are shown in Table VII. From the table, it is evident that for most distortion types, such as JPEG, WN,

TABLE VI
CROSS-DATABASE VALIDATION. THE BEST PERFORMANCES ARE IN BOLD

Train	SJTU IQSD				CSIQ			
Test	KonIQ-10K		VCLFER		KonIQ-10K		VCLFER	
Method	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE	0.1294	0.0894	0.5426	0.5501	0.3034	0.2175	0.7686	0.7574
NIQE	0.0359	0.0364	0.5764	0.5867	0.3490	0.2530	0.7365	0.7404
BMPRI	0.1378	0.1092	0.5064	0.4792	0.1268	0.1112	0.5464	0.5243
Kang	0.4028	0.3948	0.6721	0.7382	0.3865	0.3793	0.7276	0.7149
DIQaM-NR	0.4257	0.4006	0.6631	0.6962	0.4258	0.4474	0.6650	0.6582
WaDIQaM-NR	0.5750	0.5600	0.7281	0.7131	0.3184	0.3018	0.6204	0.6113
DBCNN	0.5292	0.5700	0.7931	0.7962	0.3214	0.2918	0.7411	0.7169
HyperIQA	0.5784	0.6022	0.2374	0.2411	0.3096	0.3137	0.3409	0.3346
DACNN	0.5991	0.6164	**0.8462**	0.8542	0.3686	0.3667	0.6467	0.6345
GraphIQA	0.4904	0.5108	0.2533	0.2687	0.3605	0.3722	0.4090	0.3658
NIMA	0.4499	0.4481	0.8013	0.8268	0.3887	0.4024	0.5925	0.5873
Gao	0.6253	0.6811	0.8407	0.8572	0.2388	0.2213	**0.8784**	**0.8683**
Proposed	**0.6504**	**0.6934**	0.8446	**0.8689**	**0.4934**	**0.5185**	0.7938	0.7817
Train	KonIQ-10K				CID2013			
Test	SJTU IQSD		SPAQ		SJTU IQSD		SPAQ	
Method	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE	0.0371	0.0111	0.3398	0.3368	0.4112	0.3664	0.2674	0.2182
NIQE	0.6320	0.6147	0.3098	0.3081	0.3080	0.3557	0.1999	0.2107
BMPRI	0.3998	0.3703	0.3892	0.3782	0.2829	0.1858	0.4408	0.4413
Kang	0.3991	0.4245	0.4208	0.4513	0.1198	0.1044	0.4523	0.3945
DIQaM-NR	0.2980	0.2814	0.7921	0.8025	0.1680	0.1348	0.1644	0.2128
WaDIQaM-NR	0.5457	0.4468	0.7813	0.7982	0.2761	0.2084	0.2585	0.3003
DBCNN	0.6214	0.6431	0.8035	0.8190	**0.6204**	**0.7216**	0.4915	0.4542
HyperIQA	0.6043	0.5411	0.2398	0.2287	0.5704	0.5525	0.1755	0.1799
DACNN	0.5158	0.6287	0.8167	0.8267	0.4742	0.4681	0.4909	0.4923
GraphIQA	0.7005	0.6397	0.2083	0.2598	0.6002	0.4825	0.1263	0.1227
NIMA	0.6504	0.6157	0.7737	0.7513	0.2782	0.3024	0.4613	0.5247
Gao	0.6657	0.6420	0.8198	0.8245	0.6176	0.6609	0.6901	0.6893
Proposed	**0.7251**	**0.7051**	**0.8649**	**0.8509**	0.5577	0.5625	**0.6918**	**0.6929**

TABLE VII
PREDICTION PERFORMANCE OF THE PROPOSED METHOD AND COMPETING METHODS ON INDIVIDUAL DISTORTIONS IN THE SJTU IQSD DATABASE. THE BEST PERFORMANCES ARE IN BOLD

Method	JP2K		JPEG		WN		Gblur		FF	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE	0.8507	0.8578	0.8998	0.9254	0.9195	0.9584	0.8945	0.9206	0.7401	0.8002
NIQE	0.8663	0.8848	0.8692	0.9196	0.9261	0.9577	0.8650	0.8502	0.8060	0.8214
BMPRI	0.9026	0.9151	0.9044	0.9377	0.9388	0.9617	0.8719	0.8808	0.7867	0.8210
DBCNN	0.8116	0.8128	0.8473	0.8525	0.8207	0.8937	0.8507	0.8815	0.8325	0.8704
HyperIQA	0.9436	0.9459	0.9401	0.9557	0.9656	0.9713	0.9655	0.9666	**0.9396**	0.9505
DACNN	0.9436	0.9540	0.9477	0.9217	0.9701	0.9711	0.9645	0.9654	0.9206	0.9443
GraphIQA	**0.9671**	**0.9716**	0.9487	0.9564	0.9620	0.9650	0.9438	0.9480	0.9255	0.9268
NIMA	0.6184	0.6584	0.7021	0.8124	0.7838	0.8353	0.5708	0.6300	0.5392	0.6015
Gao	0.9518	0.9592	0.9451	0.9522	0.9454	0.9740	0.9530	0.9671	0.9386	**0.9576**
StairIQA	0.9419	0.9605	0.9485	0.9568	0.9560	0.9659	0.8992	0.9186	0.8806	0.9194
Proposed	0.9575	0.9617	**0.9490**	**0.9574**	**0.9730**	**0.9741**	**0.9657**	**0.9691**	0.9346	0.9369

and Gblur, the proposed method outperforms the competing methods. For JP2K, our method achieves the second-best prediction performance. In conclusion, the proposed method demonstrates outstanding prediction performance for most individual distortions.

G. Ablation Analysis

Next, ablation analyses are conducted to demonstrate the impact and importance of different network structures in the proposed method on prediction performance.

1) Backbone Network: In this paper, the pre-trained ResNet50 is used as the backbone network of the visual feature extraction module. To evaluate the impact of different backbone networks, we replace ResNet50 with other pre-trained networks, including AlexNet [71], VGG16 [72], MobileNet [73], DenseNet [74], ViT [75], and Swin-T [76]. We compare the prediction performance of these backbone networks on the SJTU IQSD database. The comparison results are shown in Table VIII. From the table, it can be seen that the proposed method achieves the best prediction performance when ResNet50 is used as the backbone network.

2) First- and Second-Order Visual Features: After using the backbone network to extract image features, we use GAP and GMP to extract first-order visual features, while GSP

TABLE VIII

IMPACT OF DIFFERENT BACKBONE NETWORKS ON THE PREDICTION PERFORMANCE OF THE PROPOSED METHOD. THE BEST PERFORMANCES ARE IN BOLD

Backbone	SRCC	PLCC	RMSE
AlexNet	0.9169	0.9366	7.1303
VGG16	0.9370	0.9500	6.3761
MobileNet	0.9458	0.9486	6.6044
DenseNet	0.9514	0.9587	5.7814
ViT	0.8914	0.9058	8.7250
Swin-T	0.9142	0.9310	7.5359
ResNet50	**0.9622**	**0.9679**	**5.2706**

TABLE IX

IMPACT OF USING GAP, GMP, AND GSP TO EXTRACT FIRST- AND SECOND-ORDER VISUAL FEATURES ON THE PREDICTION PERFORMANCE OF THE PROPOSED METHOD. THE BEST PERFORMANCES ARE IN BOLD

Comparison Method	SRCC	PLCC	RMSE
GAP	0.9487	0.9582	5.8476
GMP	0.9404	0.9533	6.1024
GAP & GMP	0.9412	0.9489	6.3727
GSP	0.9414	0.9490	6.6445
GAP & GSP	0.9318	0.9409	6.8730
GMP & GSP	0.9314	0.9421	6.8236
GAP & GMP & GSP	**0.9622**	**0.9679**	**5.2706**

is used to extract second-order visual features. To evaluate the impact of these extracted visual features on the prediction performance of the proposed method, we construct six comparison methods: The first comparison method uses GAP to extract first-order visual features; the second comparison method uses GMP to extract first-order visual features; the third comparison method uses GAP and GMP to extract first-order visual features; the fourth comparison method uses only GSP to extract second-order visual features; the fifth comparison method uses GAP and GSP to extract first- and second-order visual features; the sixth comparison method uses GMP and GSP to extract first- and second-order visual features. The prediction performance of these six comparison methods on the SJTU IQSD database is shown in Table IX. The results indicate that the simultaneous use of GAP, GMP, and GSP to extract first- and second-order visual features can significantly improve the prediction performance of the proposed method.

3) Graph Convolutional Module: A crucial component of the proposed method is the graph convolutional module, which is used to predict the image quality score distribution. To investigate the importance of this module, we construct a comparison method that replaces the graph convolutional module with a FC layer for mapping image features to the image quality score distribution. The results are shown in Table X. From the table, it can be seen that the proposed graph convolutional module can improve the performance of the method in predicting the MOS of image quality.

4) Loss Function: This paper uses the image quality score distribution and the MOS of image quality to train the proposed method simultaneously. Accordingly, we design a distribution-based loss function and a MOS-based loss

TABLE X

IMPORTANCE OF THE GRAPH CONVOLUTIONAL MODULE FOR THE PROPOSED METHOD. THE BEST PERFORMANCES ARE IN BOLD

Method	SRCC	PLCC	RMSE
Without graph convolutional module	0.9440	0.9523	6.1812
With graph convolutional module	**0.9622**	**0.9679**	**5.2706**

TABLE XI

IMPACT OF THE DISTRIBUTION-BASED LOSS FUNCTION AND THE MOS-BASED LOSS FUNCTION ON THE PREDICTION PERFORMANCE OF THE PROPOSED METHOD. THE BEST PERFORMANCES ARE IN BOLD

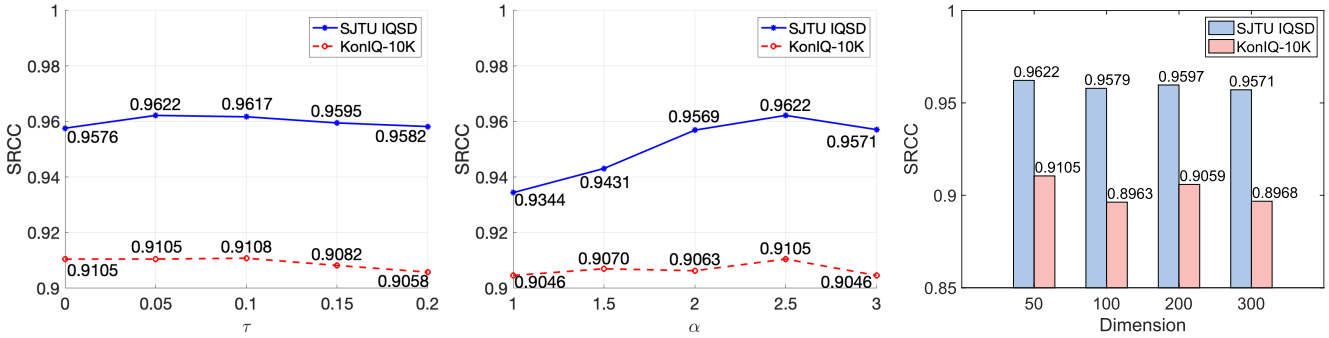
Loss Function	SRCC	PLCC	RMSE
Distribution-based	0.9387	0.9494	6.4138
MOS-based	0.9555	0.9549	7.1690
Distribution-based & MOS-based	**0.9622**	**0.9679**	**5.2706**

function. To analyze the impact of these two loss functions on the prediction performance, we construct two comparison methods. The first method uses only the distribution-based loss function for training, while the second method uses only the MOS-based loss function. The comparison results are shown in Table XI. The table shows that training the proposed method using both the MOS and the image quality score distribution effectively improves the prediction performance.

H. Parameter Analysis

1) a: To validate the effectiveness of the value of a, we train the proposed method on the SJTU IQSD, KonIQ-10K, CSIQ, LIVE MD, LIVE Challenge, and CID2013 databases using both the distribution-based loss function from Eq. (20) and the MOS-based loss function. The results are presented in Table XII. These results suggest that, even with a fixed $a = 0.1304$, the proposed method maintains outstanding prediction performance, demonstrating that this value is suitable for various categories of IQA databases. In addition, Table XII shows a relatively small decrease in prediction performance across all databases compared to Tables II and III. This indicates that using either the ground-truth image quality score distribution or the ground-truth SOS to generate the Gaussian-based image quality score distribution for training improves the prediction performance of the method. Therefore, we use the ground-truth image quality score distribution for training on the SJTU IQSD and KonIQ-10K databases, while utilizing the ground-truth SOS to generate the Gaussian-based image quality score distribution for training on the CSIQ, LIVE MD, LIVE Challenge, and CID2013 databases.

2) τ: In this paper, τ determines whether the correlation between two quality labels is significant. A higher value of τ means that only correlations with very high conditional probabilities are considered significant, while a lower value of τ allows more conditional probabilities to be considered significantly correlated. We analyze the impact of different values of τ on the prediction performance of the proposed method on the SJTU IQSD and KonIQ-10K databases, as shown in Fig. 6 (a). This figure indicates that the SRCC for both databases remains relatively stable across different values of τ,



(a) (b) (c)

Fig. 6. Analysis of τ, α, and the dimension of the GloVe model on the SJTU IQSD and KonIQ-10K databases. (a) shows the prediction performance of the proposed method with different τ. (b) shows the prediction performance of the proposed method with different α. (c) shows the prediction performance of the proposed method with different dimensions of the GloVe model.

TABLE XII

PERFORMANCE OF THE PROPOSED METHOD FOR PREDICTING THE MOS OF IMAGE QUALITY ON THE SJTU IQSD, KONIQ-10K, CSIQ, LIVE MD, LIVE CHALLENGE, AND CID2013 DATABASES USING EQ. (20) WHEN a IS SET TO 0.1304

Database	SRCC	PLCC	RMSE
SJTU IQSD	0.9583	0.9551	6.0127
KonIQ-10K	0.9063	0.9203	0.2145
CSIQ	0.9162	0.9348	0.0938
LIVE MD	0.9423	0.9450	5.9830
LIVE Challenge	0.8505	0.8715	9.6398
CID2013	0.9080	0.9191	10.1039

demonstrating that the method is robust to changes in τ. For the SJTU IQSD database, the optimal value of τ is around 0.05, where the highest SRCC is observed. As τ exceeds 0.05, the prediction performance decreases slightly. For the KonIQ-10K database, there are minimal differences in prediction performance when τ is set to 0, 0.05, and 0.1. As τ exceeds 0.1, the prediction performance decreases slightly. Taking all factors into consideration, setting τ to 0.05 may be beneficial in achieving optimal prediction performance.

3) α: The hyper-parameter α represents the weight of the MOS-based loss function. In this section, we analyze the impact of different values of α on the prediction performance of the proposed method on the SJTU IQSD and KonIQ-10K databases. As shown in Fig. 6 (b), the prediction performance on the KonIQ-10K database is more stable compared to that on the SJTU IQSD database. This stability may be attributed to the larger size of the KonIQ-10K database. When α is set to 2.5, the method achieves the best prediction performance on both the SJTU IQSD and KonIQ-10K databases. Therefore, we set the weight of the MOS-based loss function to 2.5.

4) Dimension of the GloVe Model: The GloVe model used in this paper is 50-dimensional. In addition, the authors in [42] provide 100-dimensional, 200-dimensional, and 300-dimensional GloVe models. In this section, we investigate the impact of different dimensions of GloVe models on the prediction performance of the proposed method on the SJTU IQSD and KonIQ-10K databases. The results are shown in Fig. 6 (c). From the figure, it can be seen that the dimension of the GloVe model has minimal impact on the prediction performance of the proposed method. When the 50-dimensional GloVe model is used, the prediction performance is the best. Therefore, it is appropriate to use the 50-dimensional GloVe model in this paper.

TABLE XIII

RUNNING TIME OF THE PROPOSED METHOD AND COMPETING METHODS. 'TIME' REFERS TO THE AVERAGE SECOND OF TESTING ON AN IMAGE

Method	Time (second/image)
BRISQUE	0.0152
NIQE	0.0082
BMPRI	0.1169
Kang	0.0602
DIQaM-NR	0.0645
WaDIQaM-NR	0.0606
SGDNet	0.1251
DBCNN	0.0970
HyperIQA	0.0308
DACNN	0.0629
GraphIQA	0.0319
NIMA	0.0268
Gao	0.1240
Proposed	0.1071

I. Running Time

In this section, we compare the running time of the proposed method with that of competing methods. We select 100 images from the SJTU IQSD database and resize them to 224 × 224 pixels. For network-based methods, we test the average time using a computer with the NVIDIA GeForce RTX 3090 GPU. For conventional methods, we test the average running time using the computer with the CPU. The results are shown in Table XIII. From the table, it is evident that conventional methods, such as NIQE, have the fastest running time, followed by BRISQUE. This advantage is likely due to their simpler architectures and computations. In contrast, most network-based methods have longer running times due to their more complex architectures and computations. Although the running time of our proposed method is not the shortest, it remains competitive with several competing methods. Its strong prediction performance makes it highly suitable for practical use in IQA tasks.

IV. Conclusion

In this paper, we propose a novel NR IQA method that predicts the MOS of image quality by leveraging the predicted image quality score distribution. This method is divided into three modules, including a visual feature extraction module, a graph convolutional module, and a MOS prediction module. First, we design a visual feature extraction module to extract both first- and second-order visual features from the image. Then, utilizing the correlations between quality labels, we construct a GCN-based mapper. In the MOS prediction module, this mapper is combined with the extracted image features to predict the image quality score distribution, from which the MOS can be derived. The proposed method is trained jointly using the MOS and the image quality score distribution. To address the lack of the ground-truth image quality score distribution in some IQA databases, we introduce methods to generate a Gaussian-based image quality score distribution that aligns with subjective perception and design loss functions for training. Experimental results validate the effectiveness of the proposed method and demonstrate its superior performance in predicting both the MOS of image quality and the image quality score distribution compared to state-of-the-art IQA methods.

References

[1] X. Min et al., "Screen content quality assessment: Overview, benchmark, and beyond," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–36, 2021.

[2] G. Zhai and X. Min, "Perceptual image quality assessment: A survey," *Sci. China Inf. Sci.*, vol. 63, no. 11, pp. 1–52, Nov. 2020.

[3] X. Min, J. Zhou, G. Zhai, P. L. Callet, X. Yang, and X. Guan, "A metric for light field reconstruction, compression, and display quality evaluation," *IEEE Trans. Image Process.*, vol. 29, pp. 3790–3804, 2020.

[4] X. Min, H. Duan, W. Sun, Y. Zhu, and G. Zhai, "Perceptual video quality assessment: A survey," 2024, *arXiv:2402.03413*.

[5] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, pp. 1–21, Jan. 2010.

[6] X. Min et al., "Quality evaluation of image dehazing methods using synthetic hazy images," *IEEE Trans. Multimedia*, vol. 21, no. 9, pp. 2319–2333, Sep. 2019.

[7] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *Proc. Int. Conf. Comput. Vis.*, Sep. 2020, pp. 3677–3686.

[8] X. Min, K. Gu, L. Zhang, V. Jakhetiya, and G. Zhai, "Editorial: Computational neuroscience for perceptual quality assessment," *Frontiers Neurosci.*, vol. 16, Mar. 2022, Art. no. 876969.

[9] Y. Gao et al., "VDPVE: VQA dataset for perceptual video enhancement," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2023, pp. 1474–1483.

[10] Y. Gao, X. Min, W. Zhu, X.-P. Zhang, and G. Zhai, "Image quality score distribution prediction via alpha stable model," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 6, pp. 2656–2671, Dec. 2022.

[11] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Trans. Image Process.*, vol. 29, pp. 4041–4056, 2020.

[12] Y. Tian, H. Zeng, J. Hou, J. Chen, J. Zhu, and K.-K. Ma, "A light field image quality assessment model based on symmetry and depth features," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 5, pp. 2046–2050, May 2021.

[13] S. Seo, S. Ki, and M. Kim, "A novel just-noticeable-difference-based saliency-channel attention residual network for full-reference image quality predictions," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 31, no. 7, pp. 2602–2616, Jul. 2021.

[14] G. Zhai, X. Wu, X. Yang, W. Lin, and W. Zhang, "A psychovisual quality metric in free-energy principle," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 41–52, Jan. 2012.

[15] Y. Liu, G. Zhai, K. Gu, X. Liu, D. Zhao, and W. Gao, "Reduced-reference image quality assessment in free-energy principle and sparse representation," *IEEE Trans. Multimedia*, vol. 20, no. 2, pp. 379–391, Feb. 2018.

[16] X. Min, K. Gu, G. Zhai, M. Hu, and X. Yang, "Saliency-induced reduced-reference quality index for natural scene and screen content images," *Signal Process.*, vol. 145, pp. 127–136, Apr. 2018.

[17] W. Zhang, G. Zhai, Y. Wei, X. Yang, and K. Ma, "Blind image quality assessment via vision-language correspondence: A multitask learning perspective," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Sep. 2023, pp. 14071–14081.

[18] X. Min, K. Ma, K. Gu, G. Zhai, Z. Wang, and W. Lin, "Unified blind quality assessment of compressed natural, graphic, and screen content images," *IEEE Trans. Image Process.*, vol. 26, no. 11, pp. 5462–5474, Nov. 2017.

[19] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "Generalizable no-reference image quality assessment via deep meta-learning," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1048–1060, Mar. 2022.

[20] Y. Liu et al., "Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 929–943, Apr. 2020.

[21] C. Zhang, Z. Huang, S. Liu, and J. Xiao, "Dual-channel multi-task CNN for no-reference screen content image quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 8, pp. 5011–5025, Aug. 2022.

[22] T. Song, L. Li, D. Cheng, P. Chen, and J. Wu, "Active learning-based sample selection for label-efficient blind image quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 5884–5896, Jul. 2024.

[23] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[24] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Apr. 2012.

[25] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 508–517, Jun. 2018.

[26] Y. Zhu, Y. Li, W. Sun, X. Min, G. Zhai, and X. Yang, "Blind image quality assessment via cross-view consistency," *IEEE Trans. Multimedia*, vol. 25, pp. 7607–7620, 2022.

[27] Z. Pan et al., "DACNN: Blind image quality assessment via a distortion-aware convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7518–7531, Nov. 2022.

[28] T. Song, L. Li, P. Chen, H. Liu, and J. Qian, "Blind image quality assessment for authentic distortions by intermediary enhancement and iterative training," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 11, pp. 7592–7604, Nov. 2022.

[29] T. Zhou, S. Tan, B. Zhao, and G. Yue, "Multitask deep neural network with knowledge-guided attention for blind image quality assessment," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 8, pp. 7577–7588, Aug. 2024.

[30] Z. Zhou, F. Zhou, and G. Qiu, "Blind image quality assessment based on separate representations and adaptive interaction of content and distortion," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 4, pp. 2484–2497, Apr. 2024.

[31] Z. Zhou, J. Li, D. Zhong, Y. Xu, and P. Le Callet, "Deep blind image quality assessment using dynamic neural model with dual-order statistics," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 7, pp. 6279–6290, Jul. 2024.

[32] G. Zhai and X. Zhang, "Probabilistic image quality assessment: An economics point of view," in *Proc. Int. Symp. Intell. Signal Process. Commun. Syst. (ISPACS)*, Dec. 2019, pp. 1–2.

[33] Y. Gao, X. Min, Y. Zhu, X.-P. Zhang, and G. Zhai, "Blind image quality assessment: A fuzzy neural network for opinion score distribution prediction," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 3, pp. 1641–1655, Mar. 2024.

[34] Y. Gao, X. Min, W. Zhu, X.-P. Zhang, and G. Zhai, "Parameterized image quality score distribution prediction," 2022, *arXiv:2203.00926*.

[35] H. Talebi and P. Milanfar, "NIMA: Neural image assessment," *IEEE Trans. Image Process.*, vol. 27, no. 8, pp. 3998–4011, Aug. 2018.

[36] A. Liu, J. Wang, J. Liu, and Y. Su, "Comprehensive image quality assessment via predicting the distribution of opinion score," *Multimedia Tools Appl.*, vol. 78, no. 17, pp. 24205–24222, Sep. 2019.

[37] Y. Gao, X. Min, Y. Zhu, J. Li, X.-P. Zhang, and G. Zhai, "Image quality assessment: From mean opinion score to opinion score distribution," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 997–1005.

[38] Y. Gao, X. Min, W. Zhu, X.-P. Zhang, and G. Zhai, "Modeling image quality score distribution using alpha stable model," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2021, pp. 1574–1578.

[39] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R Recommendation BT. 500-11, ITU-R, 2002.

[40] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *Proc. 46th Asilomar Conf. Signals Syst. Comput. (ASILOMAR)*, 2012, pp. 1693–1697.

[41] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Sep. 2015.

[42] J. Pennington, R. Socher, and C. Manning, "Glove: Global vectors for word representation," in *Proc. Conf. Empirical Methods Natural Lang. Process. (EMNLP)*, 2014, pp. 1532–1543.

[43] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 5177–5186.

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

[45] A. Zaric et al., "VCL@FER image quality assessment database," in *Proc. ELMAR-*, Sep. 2011, pp. 105–110.

[46] T. Hoßfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!" in *Proc. 3rd Int. Workshop Quality Multimedia Exper.*, Sep. 2011, pp. 131–136.

[47] H. Zeng, L. Zhang, and A. C. Bovik, "Blind image quality assessment with a probabilistic quality representation," in *Proc. 25th IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2018, pp. 609–613.

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

[49] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3085–3094.

[50] J. Li, Y. Han, M. Zhang, G. Li, and B. Zhang, "Multi-scale residual network model combined with global average pooling for action recognition," *Multimedia Tools Appl.*, vol. 81, no. 1, pp. 1375–1393, Jan. 2022.

[51] P. Li, J. Xie, Q. Wang, and W. Zuo, "Is second-order information helpful for large-scale visual recognition?" in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 2089–2097.

[52] J. Carreira, C. Rui, J. Batista, and C. Sminchisescu, "Semantic segmentation with second-order pooling," in *Proc. 12th Eur. Conf. Comput. Vis.*, Sep. 2012, pp. 430–443.

[53] Z. Gao, J. Xie, Q. Wang, and P. Li, "Global second-order pooling convolutional networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3019–3028.

[54] Q. Li, Z. Han, and X.-M. Wu, "Deeper insights into graph convolutional networks for semi-supervised learning," in *Proc. AAAI Conf. Artif. Intell.*, 2018, vol. 32, no. 1, pp. 3538–3545.

[55] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," 2016, *arXiv:1609.02907*.

[56] A. L. Maas, A. Y. Hannun, and A. Y. Ng, "Rectifier nonlinearities improve neural network acoustic models," in *Proc. Int. Conf. Mach. Learn.*, 2013, vol. 30, no. 1, p. 3.

[57] L. Hou, C.-P. Yu, and D. Samaras, "Squared Earth mover's distance-based loss for training deep neural networks," 2016, *arXiv:1611.05916*.

[58] D. Li, T. Jiang, and M. Jiang, "Norm-in-norm loss with faster convergence and better performance for image quality assessment," in *Proc. 28th ACM Int. Conf. Multimedia*, 2020, pp. 789–797.

[59] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*.

[60] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "ImageNet: A large-scale hierarchical image database," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Miami, FL, USA, Jun. 2009, pp. 248–255.

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

[62] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, pp. 206–219, 2017.

[63] S. Yang, Q. Jiang, W. Lin, and Y. Wang, "SGDNet: An end-to-end saliency-guided deep neural network for no-reference image quality assessment," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 1383–1391.

[64] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Jan. 2018.

[65] S. Su et al., "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Sep. 2020, pp. 3667–3676.

[66] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Uncertainty-aware blind image quality assessment in the laboratory and wild," *IEEE Trans. Image Process.*, vol. 30, pp. 3474–3486, 2021.

[67] S. Sun, T. Yu, J. Xu, W. Zhou, and Z. Chen, "GraphIQA: Learning distortion graph representations for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 25, pp. 2912–2925, 2022.

[68] W. Sun, X. Min, D. Tu, S. Ma, and G. Zhai, "Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training," *IEEE J. Sel. Topics Signal Process.*, vol. 17, no. 6, pp. 1–15, Apr. 2023.

[69] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Task-specific normalization for continual learning of blind image quality models," *IEEE Trans. Image Process.*, vol. 33, pp. 1898–1910, 2024.

[70] C. Chen et al., "TOPIQ: A top-down approach from semantics to distortions for image quality assessment," *IEEE Trans. Image Process.*, vol. 33, pp. 2404–2418, 2024.

[71] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 25, 2012, pp. 1097–1105.

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

[73] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "MobileNetV2: Inverted residuals and linear bottlenecks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 4510–4520.

[74] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *Proc. IEEE Conf. Comput Vis. Pattern Recognit.*, Aug. 2017, pp. 4700–4708.

[75] A. Dosovitskiy et al., "An image is worth 16x16 words: Transformers for image recognition at scale," 2020, *arXiv:2010.11929*.

[76] Z. Liu et al., "Swin transformer V2: Scaling up capacity and resolution," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2022, pp. 12009–12019.

Yixuan Gao received the B.E. degree from Harbin Institute of Technology, Weihai, China, in 2020. She is currently pursuing the Ph.D. degree with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University, Shanghai, China. Her current research interest include image quality assessment.

Xiongkuo Min (Member, IEEE) received the B.E. degree from Wuhan University, Wuhan, China, in 2013, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2018. From January 2016 to January 2017, he was a Visiting Student with the University of Waterloo. From June 2018 to September 2021, he was a Post-Doctoral Researcher with Shanghai Jiao Tong University. From January 2019 to January 2021, he was a Visiting Post-Doctoral Researcher with The University of Texas at Austin and the University of Macau. He is currently a tenure-track Associate Professor with the Institute of Image Communication and Network Engineering, Shanghai Jiao Tong University. His research interests include image/video/audio quality assessment, quality of experience, visual attention modeling, extended reality, and multimodal signal processing.

Yuqin Cao received the B.E. degree from Shanghai Jiao Tong University, Shanghai, China, in 2021, where she is currently pursuing the Ph.D. degree with the Institute of Image Communication and Network Engineering. Her current research interests include audio-visual quality assessment.

Xiaohong Liu (Member, IEEE) received the B.E. degree in communication engineering from Southwest Jiaotong University, Chengdu, China, in 2014, the M.A.Sc. degree in electrical and computer engineering from the University of Ottawa, Ottawa, ON, Canada, in 2016, and the Ph.D. degree in electrical and computer engineering from McMaster University, Hamilton, ON, Canada, in 2021. He is currently a tenure-track Assistant Professor with the John Hopcroft Center, Shanghai Jiao Tong University, Shanghai, China. His research interests include image/video restoration and image segmentation.

Guangtao Zhai (Senior Member, IEEE) received the B.E. and M.E. degrees from Shandong University, Shandong, China, in 2001 and 2004, respectively, and the Ph.D. degree from Shanghai Jiao Tong University, Shanghai, China, in 2009. From 2008 to 2009, he was a Visiting Student with the Department of Electrical and Computer Engineering, McMaster University, Hamilton, ON, Canada, where he was a Post-Doctoral Fellow from 2010 to 2012. From 2012 to 2013, he was a Humboldt Research Fellow with the Institute of Multimedia Communication and Signal Processing, Friedrich Alexander University Erlangen-Nuremberg, Germany. He is currently a Professor with the Department of Electronics Engineering, Shanghai Jiao Tong University. He is a member of IEEE CAS VSPC TC and MSA TC. He has received multiple international and domestic research awards, including the Best Paper Award of IEEE CVPR DynaVis Workshop 2020, the Eastern Scholar and Dawn Program Professorship of Shanghai, the NSFC Excellent Young Researcher Program, and the National Top Young Researcher Award. He is serving as the Editor-in-Chief for *Displays* (Elsevier) and is on the Editorial Board for *Digital Signal Processing* (Elsevier) and *Science China Information Sciences*.