Explainable and Generalizable Blind Image Quality Assessment via Semantic Attribute Reasoning

Yipo Huang, Leida Li, *Member, IEEE*, Yuzhe Yang, Yaqian Li, and Yandong Guo

***Abstract*—Blind image quality assessment (BIQA) that can directly evaluate image quality without perfect-quality reference has been a long-standing research topic. Although the existing BIQA models have achieved very encouraging performance, the lack of explainability and generalization ability limits their real-world applications to a great extent. People usually assess image quality according to semantic attributes, e.g., brightness, color, contrast, noise and sharpness. Furthermore, judgment on image quality is also impacted by the scene presented in the image. Therefore, the inherent relationship between semantic attributes and scenes is crucial for image quality assessment, which has rarely been explored yet. With this motivation, this paper presents a Semantic Attribute Reasoning based image QUality Evaluator (SARQUE). Specifically, we propose a two-stream network to predict semantic attributes and scene categories from distorted images. To investigate the inherent relationship between the semantic attributes and scene category, a semantic reasoning module is further proposed based on the graph convolution network (GCN), producing the final quality score. Extensive experiments conducted on five in-the-wild image quality databases demonstrate the superiority of the proposed SARQUE model over the state-of-the-arts. Furthermore, the proposed model features better explainability and generalization ability due to the use of semantic attributes.**

***Index Terms*—Blind image quality assessment, explainability, generalization ability, semantic attribute, graph convolution network**

I. Introduction

WITH the emerging popularity of smartphones and mobile internet, massive amounts of images are recorded and shared in people's daily life. In practice, images are easily contaminated by diversified distortions during their acquisition, compression, storage and transmission, which in turn cause quality degradation and impair human visual experience [1]–[4]. Objective image quality assessment (IQA), which quantifies image quality in a perceptual manner, has wide applications in image compression [5], image restoration [6], image retrieval [7], and imaging system optimization [8]–[10], etc. The current IQA models can be classified into three categories depending on the availability of perfect-quality reference images, including full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA), and no-reference or blind IQA (BIQA) [11]. Among them, BIQA models operate on distorted images directly without using any reference information, which has been the research focus recently.

Typically, a BIQA model consists of a feature extraction stage and a pooling stage [12], [13]. Early efforts mainly leverage hand-crafted features to measure the distortions, such as the mean subtracted contrast normalized (MSCN) coefficients [14], image gradient [15], and Wavelet-based features [16]. Recent works have turned to learn deep features due to the powerful representation ability of deep convolutional neural networks (CNN) [17]. During pooling, regressors, such as Support Vector Regression (SVR) [18], random forest (RF) [19], back propagation neural network (BPNN) [20], and fully connected layers (FC) [4], have been widely used to map the extracted features to an overall image quality score.

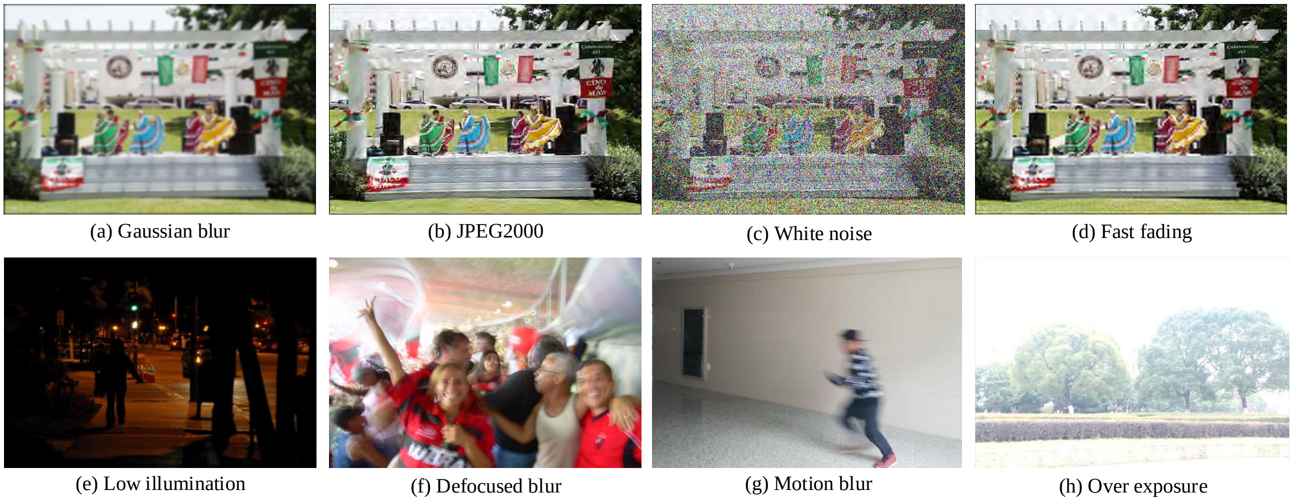
Following the above pipeline, a large number of BIQA models have been proposed in the past decade [3]. However, most of them are designed for synthetic distortions. BIQA models for authentic distortions remain extremely challenging. The underlying reasons are two-fold. (1) *Distortion diversity*. Different from the synthetic distortions that are usually generated in a lab-controlled environment, authentic distortions are much more diversified [21]. As shown in Fig. 1, synthetic distortions are usually present in the whole image homogeneously. In contrast, in-the-wild images typically suffer from a mixture of distortions, which are much more complicated to model [22]. As a result, BIQA models trained on synthetic distortions cannot easily generalize to authentic distortions. (2) *Content variation*. BIQA models are expected to have the capability of evaluating images with ever-changing content in real-world environment. However, the existing image quality databases, which are used to train BIQA models, typically consist of very limited content categories. For example, the widely used databases LIVE [23], TID2013 [24] and CSIQ [25] only contain no more than 30 content categories. Even the latest KonIQ-10K [26] and SPAQ [27] databases are only made up of about 10,000 images, which are still not representative of the complete image space. Therefore, BIQA models trained on the de facto image quality databases are still weak in handling real-world images.

In recent years, deep neural networks have demonstrated their effectiveness on BIQA [28]. Due to the lack of big training data, ImageNet [29] pre-trained model is commonly adopted to extract quality-aware features, based on which the quality score is predicted. However, image quality assessment

This work was supported in part by the National Natural Science Foundation of China under Grants 62171340, 61991451 and 61771473, the OPPO Research Fund, the Fundamental Research Funds for the Central Universities, the Innovation Fund of Xidian University under Grant YJS2202, and the Key Project of Shaanxi Provincial Department of Education (Collaborative Innovation Center) under Grant 20JY024. *(Corresponding author: Leida Li.)*

Y. Huang and L. Li are with the School of Artificial Intelligence, Xidian University, Xian 710071, China (e-mails: huangyipo@stu.xidian.edu.cn; ldli@xidian.edu.cn).

Y. Yang, Y. Li and Y. Guo are with the Intelligent Perception and Interaction Research Department, OPPO Research Institute, Shanghai, China (emails: ippllewis@gmail.com; liyaqian@oppo.com; yandong.guo@live.com).



(a) Gaussian blur

(b) JPEG2000

(c) White noise

(d) Fast fading

(e) Low illumination

(f) Defocused blur

(g) Motion blur

(h) Over exposure

Fig. 1. An example to show the difference between synthetic distortions and authentic distortions. Images (a)-(d) are degraded by synthetic distortions and images (e)-(h) are degraded by authentic distortions.

is essentially different from image recognition [4]. To be specific, BIQA models need to be sensitive to distortions, while the image recognition task is expected to be invariant to distortions. From this perspective, quality assessment and image recognition have distinct requirements for the feature representations. Although these pre-trained CNN models can mitigate the over-fitting problem to some extent, the generalization ability is quite limited. On the other hand, people tend to judge image quality according to specific semantic attributes, such as brightness, color, contrast, noise and sharpness. However, the existing BIQA models can only predict a single scalar score, which also impedes their applications in real-world scenarios. Therefore, how to design an explainable BIQA model is also of increasing concern.

Motivated by the above facts, this paper presents a Semantic Attribute Reasoning based image QUality Evaluator (SARQUE), which is characterized by good explainability and generalization ability. Specifically, a two-stream network with multiple branches is designed to learn semantic attributes and scene categories from distorted images, which are highly related to image quality. Then, we propose a semantic reasoning module based on the graph convolution network (GCN) [30] to mine the inherent relationship between semantic attributes and scene category, producing the final quality score. Extensive experimental results demonstrate that the proposed SARQUE model can not only accurately evaluate the quality of in-the-wild images, but also predict the semantic attributes simultaneously to facilitate model explainability. Further, with the help of semantic attributes, the proposed model also achieves better generalization ability. The contributions of this work can be summarized as follows.

- We propose a new BIQA model for in-the-wild images via semantic attribute reasoning, which not only delivers state-of-the-art performance, but also achieves good generalization ability. Unlike existing BIQA models that only predict a single quality score, the proposed model can also predict the quality-aware semantic attributes, which in turn facilitate better model explainability.
- We propose a GCN-based semantic reasoning module to investigate the interactions between semantic attributes and scene category in determining the overall image quality. Compared with the commonly used FC pooling, GCN-based attribute reasoning can predict image quality more comprehensively and better performance is achieved.
- We conduct extensive experiments and comparisons on five in-the-wild image quality databases, and the experimental results demonstrate the superiority of the proposed model over the state-of-the-art BIQA models. Visual results are also provided to demonstrate the explainability of the proposed model.

The rest of the paper is organized as follows. We review the related works in Section II. The details of the proposed SARQUE model are introduced in Section III. Experimental results and visual analysis are given in Section IV. Finally, conclusions are drawn in Section V.

II. Related Works

A. *Hand-crafted Feature-based BIQA*

Early works for BIQA mainly utilized hand-crafted features to measure image distortion, which can be further divided into two categories, i.e., distortion-specific metrics and general-purpose metrics. Distortion-specific metrics measure the degree of a known distortion type, such as Gaussian blur and noise [16]. Although this kind of metrics have achieved significant success, their application scope is rather limited, because the exact distortion types are usually unknown in real applications. To overcome this limitation, general-purpose BIQA metrics have been proposed. Representative general-purpose BIQA metrics include the Blind/Referenceless Image Spatial QUality Evaluator (BRISQUE) [14], BLind Image Integrity Notator using discrete cosine transform (D-CT) Statistics-II (BLIINDS-II) [31], Natural Image Quality Evaluator (NIQE) [32], Integrated Local NIQE (IL-NIQE) [33], Codebook Representation for No-Reference Image Assessment (CORNIA) [34], and BIQA method based on high

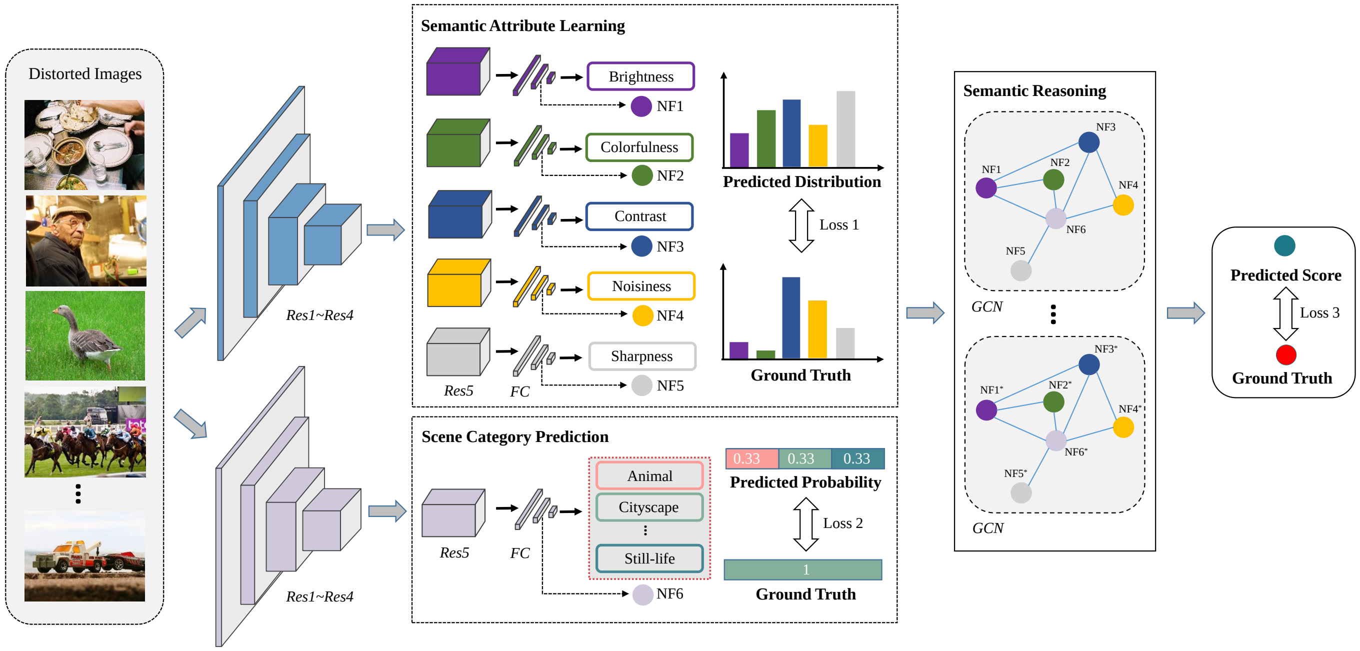


Fig. 2. The overall structure of the proposed Semantic Attribute Reasoning based image QUality Evaluator (SARQUE). Semantic Attribute Learning module learns the distribution of five semantic attributes from distorted images. Scene Category Prediction module predicts the probability of scene category. Semantic Reasoning module mines the inherent relationship between the semantic attributes and scene category. Res: block of ResNet-50; FC: fully connected layer; NF: node feature.

order statistics aggregation (HOSA) [35], just to name a few. Quality-aware features are usually extracted based on Natural Scene Statistics (NSS) or visual codebooks. This kind of features are designed based on the intuitive understanding of the distortion characteristics, so they have explicit physical meanings. However, hand-crafted features are usually not comprehensive in representing the challenging distortions of in-the-wild images.

B. Deep Learning-based BIQA

With the boom of deep learning, convolutional neural networks have become the de facto configuration in building modern BIQA models. To alleviate the conflict between the small number of training samples and the large number of learnable model parameters, early attempts mainly employed relatively shallow networks for extracting quality-aware features. For example, Kang *et al.* [36] utilized a shallow CNN model to perform quality prediction, which consists of a convolutional layer, two fully connected layers and an output node. In [37], Hou *et al.* first utilized a discriminative deep model to classify an image to five quality grades. Then, a quality pooling strategy was used to produce the final quality score. Recently, deeper networks were utilized to handle more complex distortions. In [38], Ma *et al.* proposed a novel multi-task learning framework for BIQA by learning distortion identification and quality prediction simultaneously. In [26], Vlad *et al.* built a large-scale in-the-wild IQA database and proposed a deep learning model to measure the authentic distortions. Zhang *et al.* [39] used a two-stream network to learn synthetic distortions and authentic distortions simultaneously. Then, the bi-linear pooling was adopted to predict the quality score. In [27], Fang *et al.* introduced a new authentically distorted IQA database with rich annotations and proposed a BIQA model based on multi-task learning. Liu *et al.* [40] utilized synthetically generated distortions to build a large number of image pairs, based on which a prior model was trained. Then, a target BIQA model can be easily obtained after fine tuning using a small amount of images. In [41], Ma *et al.* built an opinion-unaware BIQA model using the learning-to-rank strategy with collected large scale quality-discriminable image pairs. In [42], Zhu *et al.* leveraged meta-learning to train a general-purpose BIQA model and achieved impressive performance on both synthetic and authentic distortions.

BIQA has achieved very impressive advances, especially the deep learning-based models. However, the current BIQA models can only predict a simple scalar score. The underlying reasons why the quality score is obtained are unknown. Further, the fundamental challenges of distortion diversity and content variation lead to the generalization problem, which cannot meet the requirement of real-world applications. In this paper, we make attempts to handle the above challenges by proposing a new BIQA model via semantic attribute reasoning, which features better explainability and generalization ability.

III. Proposed Method

Fig. 2 illustrates the framework of the proposed BIQA model based on semantic attribute reasoning. The whole framework consists of three modules, namely semantic attribute learning, scene category prediction and semantic reasoning. A two-stream network is adopted to learn semantic attributes and scene category features, which are both closely related to image quality. The semantic reasoning module is built based on the GCN, which is designed to investigate the inherent relationship between semantic attributes and scene category for predicting the overall image quality score.

A. Analysis of Semantic Attributes and Scene Category

People typically assess image quality according to semantic attributes, e.g., brightness, color, contrast, noise and sharpness. Fig. 3 shows two examples of high-quality and low-quality images as well as the corresponding mean opinion score (MOS) values and semantic attributes. The high-quality image has good brightness, vivid color and high sharpness. In contrast, the low-quality image has poor brightness, dull color and low sharpness. Therefore, semantic attributes strongly correlate with image quality, which can describe the quality of an image intuitively. Furthermore, the scene category presented in an image impact people's judgment on image quality. For example, we tend to regard an image of a clear blue sky as having high quality, while for quality prediction models, it is most likely to be regarded as blur-contaminated due to the large homogeneous area [1]. This is mainly because that human can differentiate the scene categories when judging image quality [43]. However, content variation is still an open challenge in BIQA.

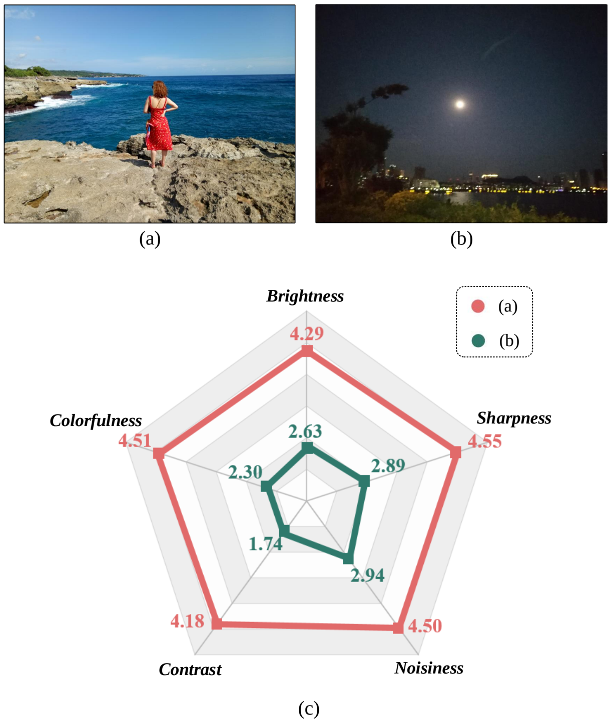
Both semantic attributes and scene categories are crucial for IQA, and they are related to each other. Therefore, there is a strong correlation between semantic attributes and scene categories, which interact to produce the overall quality score. Inspired by the above facts, we first propose a two-stream network to learn semantic attributes and scene category, respectively. Then, we design a semantic reasoning model based on GCN to mine the inherent connections between the two components for predicting the final quality score.

B. Semantic Attribute Learning

Attribute learning, which enables semantic expression of features in deep models, is frequently used in computer vision [44]. For example, in [45], the authors found that visual attributes benefit the learning of effective image representations, achieving superior performance on object recognition task. In [46], the authors adopted attribute learning in zero-shot classification and achieved higher classification accuracy. As aforementioned, the semantic attributes are highly related to image quality, such as brightness, colorfulness, contrast, noisiness and sharpness. In this part, we use a multi-branch CNN model to learn the quality-aware semantic attributes, which is shown in Fig. 2. Specifically, we implement the network of semantic attribute learning using the five blocks (denoted as *Res1* - *Res5*) of the ResNet-50 backbone [17], where *Res1* - *Res5* represent conv1, conv2_x, conv3_x, conv4_x and conv5_x, respectively. In this work, the first four blocks (*Res1* - *Res4*) are used as the shared feature extraction module. For an input image x, the hidden features $\boldsymbol{d}_a$ are obtained from the shared feature extraction module F_{θ_a} as:

$$\boldsymbol{d}_a = F_{\theta_a}(x), \tag{1}$$

where θ_a denotes the parameter set of the shared feature extraction module F_{θ_a}.


(a)

(b)

(c)

Fig. 3. Examples of high-quality and low-quality images as well as the corresponding semantic attributes. (a) high-quality image (MOS=4.49); (b) low-quality image (MOS=2.21); (c) attribute distribution map of (a) and (b).

Then, we add another four residual blocks that are exactly the same as the residual block *Res5*, and the five parallel residual blocks are used to predict the five semantic attributes respectively. Following each of the five branches, we add three FC layers with PReLU activation function, which contain 256 nodes, 64 nodes and 1 node, respectively. Next, we leverage the five attribute branches to further map the hidden features $\boldsymbol{d}_a$ to the semantic attributes $\hat{\boldsymbol{o}}$, which is defined as:

$$\hat{\boldsymbol{o}} = F_{\theta_k}(\boldsymbol{d}_a), \tag{2}$$

where θ_k denotes the parameters of each attribute branch F_{θ_k}, and $\hat{\boldsymbol{o}} = \{\hat{o}_1, \hat{o}_2, \ldots, \hat{o}_k\}$ denotes the semantic attributes.

During semantic attribute learning, we assume that $\mathcal{D} = \{x_i, \boldsymbol{o}_i, \boldsymbol{s}_i, \boldsymbol{q}_i\}_{i=1}^{N_a}$ can provide images and corresponding semantic attribute labels, where $\boldsymbol{o}_i$ denotes the labeled semantic attributes of image x_i $(i = 1, 2, 3, \ldots, N_a)$. Based on the dataset, l_1 loss is leveraged to optimize the parameters θ_a and θ_k, which is defined as:

$$\mathcal{L}_1 = \frac{1}{N_a} \sum_{i=1}^{N_a} \mid \boldsymbol{o}_i - \hat{\boldsymbol{o}}_i \mid, \tag{3}$$

where $\hat{\boldsymbol{o}}_i$ denote the predicted semantic attributes of image x_i, which is computed by:

$$\hat{\boldsymbol{o}}_i = F_{\theta_k}(F_{\theta_a}(x_i)). \tag{4}$$

Following the above steps, the semantic attribute learning module can be built by training on $\mathcal{D}$ and can simultaneously

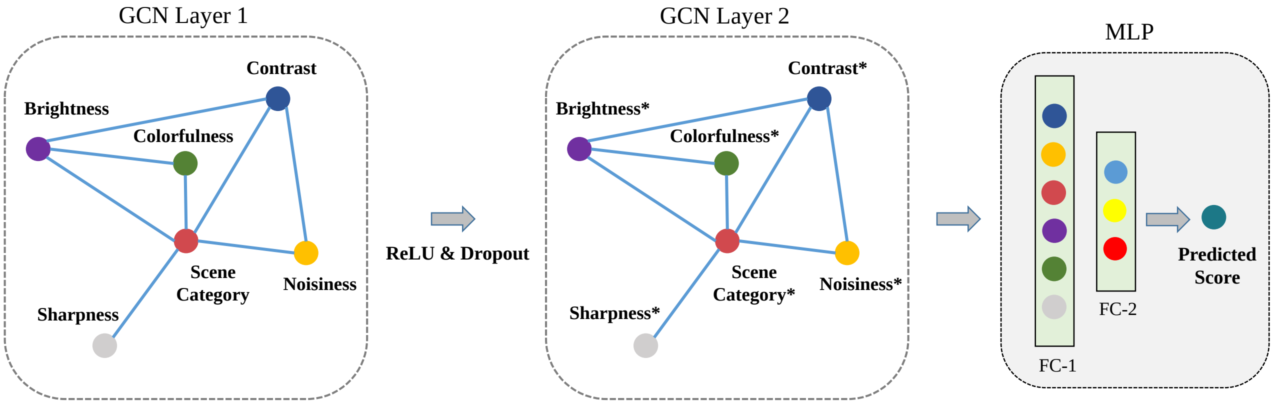


Fig. 4. Network architecture of the GCN-based semantic reasoning module.

extract the features of the semantic attributes, which in turn are used as inputs of the subsequent semantic reasoning module. The semantic attributes obtained here provide intuitive explanation on how the quality score is predicted.

C. Scene Category Prediction

In [43], the authors have experimentally demonstrated that building an IQA model without considering image content can only achieve sub-optimal results. The underlying reason is that human judgment on image quality is coupled with the understanding of the visual content presented. However, as aforementioned, content variation is a fundamental challenge in IQA [1]. As a result, when using semantic attributes for building the BIQA model, the scene category of the image should be considered to achieve comprehensive prediction. To this end, a scene category prediction module is introduced in the proposed model as shown in Fig. 2. Specifically, the five blocks (*Res1* - *Res5*) of another ResNet-50 backbone [17] is utilized for feature extraction. Then, we append three FC layers with PReLU activation function to map the input image x to the probability of the predicted scene categories $\hat{\boldsymbol{s}}$, which can be formulated as:

$$\hat{\boldsymbol{s}} = F_{\theta_s}(x), \tag{5}$$

where θ_s denotes the parameters of scene category prediction module F_{θ_s}, $\hat{\boldsymbol{s}} = \{\hat{\boldsymbol{s}}_1, \hat{\boldsymbol{s}}_2, \ldots, \hat{\boldsymbol{s}}_n\}$, and n is the number of scene categories.

In this module, we use $\mathcal{D} = \{x_i, \boldsymbol{o}_i, \boldsymbol{s}_i, \boldsymbol{q}_i\}_{i=1}^{N_a}$ to provide images and corresponding scene category labels, where $\boldsymbol{s}_i$ denotes the labeled scene categories of image x_i $(i = 1, 2, 3, \ldots, N_a)$. Then, l_1 loss is also adopted to optimize the parameters θ_s, which is defined as:

$$\mathcal{L}_2 = \frac{1}{N_a} \sum_{i=1}^{N_a} \mid \boldsymbol{s}_i - \hat{\boldsymbol{s}}_i \mid, \tag{6}$$

where $\hat{\boldsymbol{s}}_i$ denotes the predicted scene category of image x_i, which is computed by:

$$\hat{\boldsymbol{s}}_i = F_{\theta_s}(x_i). \tag{7}$$

In this way, the proposed scene category prediction module is trained on $\mathcal{D}$, which can not only obtain the scene category in the image, but also generate the scene features, which will be also input to the subsequent semantic reasoning module.

D. Semantic Reasoning

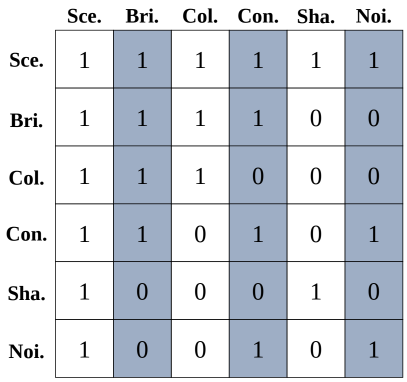
The motivation of this work is to perform IQA by mining the relationship between semantic attributes and scene category. However, the five sets of semantic attribute feature and one set of scene category feature have specific meanings, and they have different connection relationships in the IQA process. Therefore, these different feature sets belong to non-Euclidean data, directly using MLP for quality regression may lead to suboptimal results(as shown in Tables VIII and IX. Considering that the GCN model can be used to model the relationship between different nodes, in this part, we design a GCN-based semantic reasoning module to mine the inherent relationship between the semantic attributes and scene category for jointly determining the final image quality score. The framework of the proposed semantic reasoning module is depicted in Fig. 4. Inspired by the capability of relationship reasoning of GCN [47], we introduce a two-layer GCN model to perform the semantic reasoning. Different from the standard convolutions that compute on Euclidean structures in an image, the idea of GCN is to learn a mapping function $F_{\theta_g}(\cdot, \cdot)$ on a graph $\mathcal{G}$. For $F_{\theta_g}(\cdot, \cdot)$, the inputs are the feature representations $\boldsymbol{H}^l \in \mathbb{R}^{n \times d}$ and the corresponding adjacency matrix $\hat{\boldsymbol{M}} \in \mathbb{R}^{n \times n}$, where n represents the number of nodes, d denotes the dimensionality of node features, and the node features will be updated as $\boldsymbol{H}^{l+1} \in \mathbb{R}^{n \times d'}$. Therefore, a GCN layer can be formulated as:

$$\boldsymbol{H}^{l+1} = F_{\theta_g}(\boldsymbol{H}^l, \hat{\boldsymbol{M}}). \tag{8}$$

By performing the graph convolutional operation [30], $F_{\theta_g}(\cdot, \cdot)$ can be described as:

$$\boldsymbol{H}^{l+1} = h(\hat{\boldsymbol{D}}^{-\frac{1}{2}} \hat{\boldsymbol{M}} \hat{\boldsymbol{D}}^{-\frac{1}{2}} \boldsymbol{H}^l \boldsymbol{W}^l), \tag{9}$$

where $\boldsymbol{W}^l \in \mathbb{R}^{d \times d'}$ denotes the layer-specific trainable weight matrix, and $h(\cdot)$ denotes a non-linear mapping, which is



	Sce.	Bri.	Col.	Con.	Sha.	Noi.
Sce.	1	1	1	1	1	1
Bri.	1	1	1	1	0	0
Col.	1	1	1	0	0	0
Con.	1	1	0	1	0	1
Sha.	1	0	0	0	1	0
Noi.	1	0	0	1	0	1

Fig. 5. Illustration of the adjacency matrix $\widehat{\boldsymbol{M}}$. Sce.: Scene category; Bri.: Brightness; Col.: Colorfulness; Con.: Contrast; Sha.: Sharpness; Noi.: Noisiness.

achieved using ReLU. $\widehat{\boldsymbol{M}}$ is the is the adjacency matrix of the graph $\mathcal{G}$ with added self-connections,

$$\widehat{\boldsymbol{M}} = \boldsymbol{M} + \boldsymbol{E}, \tag{10}$$

where $\boldsymbol{M}$ denotes the normalized version of the graph $\mathcal{G}$.

In this work, we build the graph $\mathcal{G}$ based on the predefined manner by considering two perspectives, including 1) considering the relationship between semantic attributes and scene categories, we take scene category feature as the central node, and the five attribute nodes are all connected to scene category nodes; 2) considering the relationship between different semantic attributes, we connect brightness and color, brightness and contrast, contrast and noise, respectively. Based on this, the adjacency matrix $\widehat{\boldsymbol{M}}$ can be calculated as shown in Fig. 5. By this means, the final feature representation $\boldsymbol{H}^*$ is obtained by updating the two GCN layers:

$$\boldsymbol{H}^* = F_{\theta_{g2}}\left(F_{\theta_{g1}}(\boldsymbol{H}^l, \hat{\boldsymbol{M}}), \hat{\boldsymbol{M}}\right), \tag{11}$$

where θ_{g1} and θ_{g2} denote the parameters of first GCN layer $F_{\theta_{g1}}$ and second GCN layer $F_{\theta_{g2}}$ respectively, and $\boldsymbol{H}^l \in \mathbb{R}^{n\times d}$ is composed of the corresponding output features of the first fully connected layer in the semantic attribute learning module and the scene category prediction module. Here, we set $n = 6$ and $d = 256$. Finally, we convert the feature representation $\boldsymbol{H}^*$ into a feature vector $\mathbf{x}^*$, and use a multi-layer perceptron MLP_{θ_m} to produce the overall quality score $\hat{\boldsymbol{q}}$, which is defined as:

$$\hat{\boldsymbol{q}} = g^{(m)}(\sum_{i=0}^{L} \mathbf{x}_i^* \boldsymbol{w}_i^*), \tag{12}$$

where $\boldsymbol{w}_i^*$ represents the trainable weight of MLP_{θ_m}, $g^{(m)}$ represents the activation function, and m denotes the number of linear layers. In this work, the MLP consists of two linear layers with 8 and 1 nodes respectively. The detailed network architecture of the GCN-based semantic reasoning module is shown in Fig. 4.

In this module, we use $\mathcal{D} = \{x_i, \boldsymbol{o}_i, \boldsymbol{s}_i, \boldsymbol{q}_i\}_{i=1}^{N_a}$ to denote the image quality dataset, where $\boldsymbol{q}_i$ denotes the ground truth quality score of image x_i $(i = 1, 2, 3, \ldots, N_a)$. Based on this dataset, we compute the l_1 loss to optimize the parameters of the whole model, which is defined as:

$$\mathcal{L}_3 = \frac{1}{N_a}\sum_{i=1}^{N_a} \mid \boldsymbol{q}_i - \hat{\boldsymbol{q}}_i \mid . \tag{13}$$

Algorithm 1 The proposed SARQUE model.

```
Input: IQA dataset $\mathcal{D}$, which consists of three subsets including semantic attribute subset $\mathcal{D}_{attr} = \{x_i, \boldsymbol{o}_i\}_{i=1}^{N_a}$, scene category subset $\mathcal{D}_{sce} = \{x_i, \boldsymbol{s}_i\}_{i=1}^{N_a}$ and quality score subset $\mathcal{D}_{qua} = \{x_i, \boldsymbol{q}_i\}_{i=1}^{N_a}$.
Output: Predicted quality score $\hat{\boldsymbol{q}}_t$, predicted semantic attributes $\{\hat{\boldsymbol{o}}\}_{i=1}^{k}$ and predicted scene categories $\{\hat{\boldsymbol{s}}\}_{i=1}^{k}$;
Initialize all the parameters of the proposed model;
// Semantic Attribute Learning ;
For iteration $= 1, 2, \ldots,$ do;
    Sample a batch of $k$ images from $\mathcal{D}_{attr}$;
    For $j = 1, 2, \ldots, N$ do;
        Output semantic attributes $\{\hat{\boldsymbol{o}}\}_{i=1}^{k}$ by using $F_{\theta_a}$ and $F_{\theta_k}$ ;
        Update parameter $\theta_a$ and $\theta_k$ with $\mathcal{L}_1$;
    end For
end For
// Scene Category Prediction ;
For iteration $= 1, 2, \ldots,$ do;
    Sample a batch of $k$ images from $\mathcal{D}_{sce}$;
    For $j = 1, 2, \ldots, N$ do;
        Output scene categories $\{\hat{\boldsymbol{s}}\}_{i=1}^{k}$ by using $F_{\theta_s}$;
        Update parameter $\theta_s$ by computing $\mathcal{L}_2$;
    end For
end For
// Semantic Reasoning ;
Building the adjacency matrix $\boldsymbol{M}$ ;
For iteration $= 1, 2, \ldots,$ do;
    Sample a batch of $k$ images from $\mathcal{D}_{qua}$;
    For $j = 1, 2, \ldots, N$ do;
        Output quality score $\{\hat{\boldsymbol{q}}\}_{i=1}^{k}$ by using $F_{\theta_{g1}}$, $F_{\theta_{g2}}$ and $MLP_{\theta_m}$;
        Update all parameters with $\mathcal{L}_3$;
    end For
end For
Input test image $x_t$ into the trained SARQUE model;
return Predicted quality score $\hat{\boldsymbol{q}}_t$, predicted semantic attributes $\{\hat{\boldsymbol{o}}\}_{i=1}^{k}$ and predicted scene categories $\{\hat{\boldsymbol{s}}\}_{i=1}^{k}$;
```

In this work, $\mathcal{L}_1$ and $\mathcal{L}_2$ are first employed for optimizing the semantic attribute learning module and scene category prediction module on the SPAQ [27] database, respectively. Then, $\mathcal{L}_3$ is used to train the whole model on target databases, producing the final quality scores. The training process of the proposed SARQUE model is summarized in Algorithm 1.

TABLE I

SUMMARY OF IN-THE-WILD IMAGE QUALITY DATABASES WITH RESPECT TO NUMBERS OF IMAGES (# IMAGES), NUMBERS OF CAMERAS (# CAMERAS), SUBJECTIVE ENVIRONMENT, NUMBERS OF ATTRIBUTES (# ATTRIBUTES), NUMBERS OF SCENE CATEGORIES (# SCENE CATEGORIES), AND SCORE RANGE, WHERE HIGHER SCORE INDICATES BETTER QUALITY.

Database	# Images	# Cameras	Subjective environment	# Attributes	# Scene categories	Score range
SPAQ [27]	11,125	66	Laboratory	5	9	0-100
KonIQ-10k [26]	10,073	N/A	Crowdsourcing	4	N/A	1-5
LIVEW [48]	1,162	15	Crowdsourcing	N/A	N/A	0-100
RBID [49]	585	1	Laboratory	N/A	N/A	0-5
CID2013 [50]	480	79	Laboratory	4	N/A	0-100

IV. EXPERIMENTS AND ANALYSIS

A. Databases

To verify the performance of the proposed SARQUE model, we conduct a series of experiments on five in-the-wild image quality databases, including SPAQ [27], KonIQ-10k [26], LIVEW [48], RBID [49] and CID2013 [50].

SPAQ [27]. This database contains a total of 11,125 images captured by 66 smartphones. These images are degraded by authentic camera distortions, e.g., out-of-focus blurring, motion blurring, contrast reduction, and over-exposure, etc. It is worth mentioning that the SPAQ database has rich annotations. In addition to the overall MOS, it also provides labels for five semantic attributes (e.g., brightness, color, contrast, noise and sharpness) and nine scene categories.

KonIQ-10k [26]. It contains 10,073 authentically distorted images selected from the YFCC100M [58] database, and each image has more than 120 ratings. Images in this database have a wide and uniform distribution of content, brightness and sharpness. The MOS values collected by crowdsourcing are utilized as the ground truth.

LIVEW [48]. This database consists of 1,162 images with authentic distortions. Similar to SPAQ, the distorted images are captured by mobile devices. The quality score of each image is collected from 8,500 people by crowdsourcing.

RBID [49]. It has 585 images with authentic blur distortion, including out-of-focus blur, simple motion blur, and complex motion blur. The quality of each image is reported in the form of MOS with the range of [0, 5].

CID2013 [50]. This database consists of 480 images acquired by 79 digital cameras and the corresponding quality scores. The quality scores range from 0 to 100.

For clarity, detailed information of all the image quality databases is summarized in Table I.

B. Implementation Details

For all experiments, we first resize images into $244\times244\times3$, then we randomly crop them to $224\times224\times3$ with a randomly horizontal flip to augment training images. In the test stage, we directly resize original images into $224\times224\times3$ and predict the quality scores. In implementation, we first employ the SPAQ database to pre-train the semantic attribute learning and scene category prediction modules. Since semantic attribute learning and scene category prediction belong to different vision tasks, we train two ResNet-50 networks respectively. Then, target databases are used to fine-tune the entire network for the quality assessment. Specifically, the stochastic gradient descent (SGD) is used as the optimizer, and the initial learning rate is 0.03 with a warm-up strategy. After pre-training the semantic attribute learning and scene category prediction modules, we train the semantic reasoning module for 20 epochs and then adopt a warm-up strategy to train the whole network. When the loss does not decline for 20 epochs, the learning rate drops by a factor of 0.3. When the learning rate is smaller than 1×10^{-5}, we terminate the training process. We utilize Pytorch to implement the proposed model. The model is trained using a computer with Intel Core i7-9700K CPU @ 3.60GHz, and NVIDIA GeForce RTX 3090 24G GPU.

We adopt two widely used criteria for performance evaluation, including Pearson linear correlation coefficient (PLCC) and Spearman rank order correlation coefficient (SRCC). PLCC is used to evaluate the prediction accuracy, and SRCC is used to evaluate the prediction monotonicity. A better quality metric should achieve higher PLCC and SRCC values. To compute PLCC, the following five-parameter nonlinear mapping is first performed:

$$f(x)=\xi_1\left(0.5-\frac{1}{1+e^{\xi_2(x-\xi_3)}}\right)+\xi_4x+\xi_5, \tag{14}$$

where x donates the prediction score, $f(x)$ denotes the mapped score, and ξ_i, i=1, 2, …, 5, are the fitting parameters. Then, PLCC is computed based on the ground truth values and the corresponding predicted values after the five-parameter nonlinear mapping, which is defined as:

$$PLCC=\frac{\sum_{i=1}^{n}(z_i-\bar{z})(p_i-\bar{p})}{\sqrt{\sum_{i=1}^{n}(z_i-\bar{z})^2}\sqrt{\sum_{i=1}^{n}(p_i-\bar{p})^2}}, \tag{15}$$

where n denotes the number of distorted images in the database, z_i and p_i denote the ground truth value and predicted value of the ith image, and $\bar{z}$ and $\bar{p}$ denote the corresponding average values of all images. SRCC is defined as:

$$SRCC=1-\frac{6\sum_{i=1}^{n}d_i^2}{n(n^2-1)}, \tag{16}$$

where d_i denotes the difference between the ith image's ranks in the subjective and objective evaluations.

C. Performance Evaluation

We first compare the performance of the proposed SARQUE model with the relevant state-of-the-arts. Specifically, six representative hand-crafted feature-based methods are

TABLE II
Performance comparison between the proposed SARQUE model and the state-of-the-art methods on five image quality databases: SPAQ [27], KonIQ-10k [26], LIVEW [48], CID2013 [50], and RBID [49]. Weighted average PLCC/SRCC values are computed by considering the number of images in each database, i.e., bigger databases are assigned bigger weights, and results with * are obtained from published papers.

Metric	SPAQ [27]		KonIQ-10k [26]		LIVEW [48]		CID2013 [50]		RBID [49]		Weighted Average	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
BLIINDS-II [31]	0.678	0.665	0.615	0.529	0.507	0.463	0.565	0.487	0.495	0.428	0.636	0.587
BRISQUE [14]	0.832	0.822	0.689	0.647	0.645	0.607	0.648	0.615	0.617	0.594	0.752	0.726
IL-NIQE [33]	0.705	0.687	0.537	0.501	0.589	0.594	0.538	0.346	0.435	0.390	0.617	0.588
NFERM [51]	0.832	0.823	0.725	0.689	0.562	0.517	0.825	0.823	0.585	0.559	0.767	0.745
CORNIA [34]	0.867	0.859	0.773	0.738	0.662	0.618	0.680	0.624	0.712	0.695	0.809	0.786
HOSA [35]	0.873	0.866	0.791	0.761	0.678	0.659	0.685	0.663	0.716	0.684	0.820	0.802
deepIQA [52]	/	/	0.606*	0.604*	0.482*	0.493*	/	/	/	/	0.593	0.593
BIECON [53]	/	/	/	/	0.613*	0.595*	0.620*	0.606*	/	/	0.615	0.599
MEON [38]	/	/	/	/	0.693*	0.688*	0.703*	0.701*	/	/	0.696	0.692
WaDIQaM-NR [54]	/	/	0.761*	0.739*	0.680*	0.671*	0.729*	0.708*	0.742*	0.725*	0.751	0.731
DistNet-Q3 [55]	/	/	0.710*	0.702*	0.601*	0.570*	/	/	/	/	0.699	0.688
DIQA [56]	/	/	/	/	0.704*	0.703*	0.720*	0.708*	/	/	0.709	0.705
NSSADNN [57]	/	/	/	/	0.813*	0.745*	0.825*	0.748*	/	/	0.817	0.746
HyperNet [21]	0.914	0.909	0.917*	**0.906***	**0.882***	**0.859***	/	/	**0.878***	**0.869***	0.913	0.904
DB-CNN [39]	0.915*	0.911*	0.892	0.868	0.869*	0.851*	0.871	0.863	0.859*	0.845*	0.901	0.887
MetaIQA [28]	0.871	0.870	0.887*	0.850*	0.835*	0.802*	0.784*	0.766*	0.777	0.746	0.872	0.853
SARQUE (Proposed)	**0.922**	**0.918**	**0.923**	0.901	0.873	0.855	**0.934**	**0.930**	0.861	0.846	**0.919**	**0.906**

TABLE III
Performance of semantic attribute learning and scene category learning .

Brightness		Colorfulness		Contrast		Sharpness		Noisiness		Scene category
PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	Accuracy
0.840	0.818	0.814	0.801	0.819	0.815	0.913	0.904	0.831	0.833	85.3%

compared, including BLIINDS-II [31], BRISQUE [14], IL-NIQE [33], NFERM [51], CORNIA [34] and HOSA [35]. Ten top-performing deep learning-based BIQA models are also compared, including deepIQA [52], BIECON [53], MEON [38], WaDIQaM-NR [54], DistNet-Q3 [55], DIQA [56], NSSADNN [57], HyperNet [21], DB-CNN [39], and MetaIQA [28]. Following the commonly experimental setting in BIQA [21], [28], [59], we train all models with 80% randomly selected images of a dataset and test on the rest 20% images. For each train-test splitting, all models use the same training and test sets. To avoid bias, we repeat this procedure 10 times and report the median PLCC and SRCC values. All experimental results are summarized in Table II, where we highlight the best results in boldface while the second-best results are underlined.

It is known from Table II that the proposed SARQUE model delivers the top two performances on all databases. Particularly, on the SPAQ database, the proposed model is advantageous over all the compared metrics in terms of both prediction accuracy and monotonicity. On the KonIQ-10k database, SARQUE has the best prediction accuracy, and the SRCC value ranks the second, which is competitive to HyperNet [21]. On the CID2013 database, our model achieves the best performance on prediction accuracy and monotonicity. On the LIVEW and RBID databases, our model delivers the second-best prediction accuracy and monotonicity (only slightly worse than HyperNet [21]). In summary, we can observe from the weighted average PLCC/SRCC values that SARQUE achieves the best overall performance in evaluating the quality of images with authentic distortions.

In this work, we train the semantic attribute learning module and the scene category prediction module, which also provides a means to intuitively understand the underlying reasons why a specific quality score is predicted. To evaluate the performances of semantic attribute learning module and the scene category prediction module, we test the prediction accuracy of these two modules. In implementation, PLCC and SRCC are employed for evaluating the performance of the semantic attribute learning module. The prediction accuracy is adopted to evaluate the performance of scene recognition, which represents the proportion of scenes that are correctly identified. The experimental results are summarized in Table III. From the table, we know that the prediction of five semantic attributes achieves very encouraging results, and the PLCC and SRCC values are all higher than 0.8. Especially for sharpness, the PLCC and SRCC values are both higher than 0.9. Furthermore, the prediction accuracy of the scene category reaches 85.3%. The above experimental results show that our model achieves very encouraging performance in predicting semantic attributes and scene categories.

TABLE IV
COMPARISON RESULT OF THE PROPOSED SARQUE MODEL WITH SIX STATE-OF-THE-ART BIQA METRICS BY TRAINING ON SPAQ DATABASE AND DIRECTLY TESTING ON OTHER DATABASES: KONIQ-10K [26], LIVEW [48], CID2013 [50], AND RBID [49].

Metric	KonIQ-10k [26]		LIVEW [48]		CID2013 [50]		RBID [49]	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
BRISQUE [14]	0.446	0.433	0.593	0.553	0.499	0.504	0.589	0.578
NFERM [51]	0.455	0.447	0.591	0.542	0.437	0.342	0.578	0.570
CORNIA [34]	0.532	0.516	0.663	0.621	0.552	0.465	0.676	0.673
HOSA [35]	0.559	0.534	0.682	0.650	0.593	0.536	0.681	0.670
MT-S [27]	0.486	0.485	0.539	0.493	0.342	0.389	0.530	0.529
HyperNet [21]	0.679	0.645	0.695	0.680	0.624	0.585	0.648	0.647
MetaIQA [28]	0.722	0.686	0.765	0.731	0.737	0.695	0.743	0.735
w/o semantic attribute reasoning	0.767	0.726	0.765	0.743	0.612	0.514	0.750	0.749
SARQUE (Proposed)	**0.803**	**0.778**	**0.791**	**0.780**	**0.740**	**0.701**	**0.776**	**0.769**

TABLE V
PERFORMANCE OF OUR SARQUE MODEL EQUIPPED WITH DIFFERENT COMPONENTS ON TARGET DATABASES: SPAQ [27], KONIQ-10K [26], LIVEW [48], CID2013 [50], AND RBID [49].

Metric	SPAQ [27]		KonIQ-10k [26]		LIVEW [48]		CID2013 [50]		RBID [49]	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Attribute	0.915	0.912	0.915	0.896	0.845	0.810	0.841	0.861	0.837	0.792
Attribute+Scene	0.919	0.915	0.918	0.898	0.852	0.824	0.870	0.875	0.845	0.833
Attribute+Scene+Reasoning	**0.922**	**0.918**	**0.923**	**0.901**	**0.873**	**0.855**	**0.934**	**0.930**	**0.861**	**0.846**

D. Generalization Ability

Considering that no validation sets may produce the risk of overfitting the test sets, we further conduct cross-dataset tests to verify the generalizable performance of the proposed SARQUE model. Specifically, we train the proposed model on the SPAQ database and then *directly* test it on other databases without doing any fine-tuning. Since most of the existing BIQA models did not report such experimental results, for fair comparison, we select several top-performing BIQA models with open source codes, and conduct the experiments under the same setting. The compared metrics include BRISQUE [14], NFERM [51], CORNIA [34], HOSA [35], MT-S [27], HyperNet [21], and MetaIQA [28]. In addition, to verify the contribution of semantic attributes on the generalization ability of the proposed SARQUE model, we perform the ablation experiments by removing semantic attribute learning and replacing GCN with FC layers (denoted as w/o semantic attribute reasoning). The experimental results are listed in Table IV.

From Table IV, we can observe that the proposed SARQUE model surpasses all the competing BIQA models with significant margins for both PLCC and SRCC values on all databases. Especially on the KonIQ-10k database, SARQUE obtains performance gains of 8.1% in terms of PLCC and 9.2% in terms of SRCC beyond MetaIQA [28]. From these results, it is evident that SARQUE delivers the best generalization performance. Moreover, when the proposed model is trained without semantic attribute reasoning, the generalization ability is degraded on all databases. A possible reason is that people's judgment of image quality often depends on the perception of semantic attributes, so semantic attributes are the important characteristics of image quality for any kind of distortions. In other words, semantic attributes are the general judgment basis for perceptual image distortion. In this work, these semantic attributes can be regarded as middle-level quality features, which can get rid of the constraints of limited distortion in the existing databases to a certain extent, and improve the model generalization ability.

E. Ablation study

To explore the effectiveness of components in the proposed SARQUE model, ablation studies are further conducted. In this experiment, we first examine the effectiveness of the sematic attribute learning module by merging the features of the five attribute branches and then use four FC layers to generate image quality score (denoted as attribute). Then, we demonstrate the performance of SARQUE when using both sematic attribute learning and scene category prediction module, but removing the GCN model and replacing it with four FC layers to generate image quality score (denoted as attribute+scene). Finally, three components of our SARQUE model are used jointly to predict image quality score (denoted as attribute+scene+reasoning). The experimental results are listed in Table V, where the best results for each database are shown boldfaced.

It is observed from Table V that when the sematic attribute learning module is used, the performance on all databases are already better than most state-of-the-art BIQA models, which can be seen from Table II. In addition, when the two components of the SARQUE model are combined to train and test on target databases, the performance further improves. Further, the best performance is achieved by using three components. On the other hand, the performance improvement

TABLE VI

Performance of our SARQUE model equipped with different pre-trained weights on target databases: SPAQ [27], KonIQ-10k [26], LIVEW [48], CID2013 [50], and RBID [49].

Metric	SPAQ [27]		KonIQ-10k [26]		LIVEW [48]		CID2013 [50]		RBID [49]	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
w/o Pre-train	0.865	0.859	0.814	0.791	0.482	0.417	0.606	0.524	0.318	0.316
w/ ImageNet Pre-train	0.909	0.904	0.875	0.856	0.819	0.794	0.911	0.907	0.819	0.786
w/o Attributes Pre-train	0.915	0.911	0.909	0.883	0.823	0.804	0.921	0.919	0.841	0.793
w/ Attribute+Scene Pre-train	**0.922**	**0.918**	**0.923**	**0.901**	**0.873**	**0.855**	**0.934**	**0.930**	**0.861**	**0.846**

TABLE VII

Performance of the proposed model per-trained on a single semantic attribute and test on target databases: SPAQ [27], KonIQ-10k [26], LIVEW [48], CID2013 [50], and RBID [49]. The weighted average PLCC/SRCC values are computed by considering the number of images in each database.

Metric	SPAQ [27]		KonIQ-10k [26]		LIVEW [48]		CID2013 [50]		RBID [49]		Weighted Average	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
w/ Brightness Pre-train	0.918	0.914	0.912	0.893	0.822	0.807	0.849	0.865	0.854	0.817	0.908	0.896
w/ Colorfulness Pre-train	0.913	0.910	0.918	0.896	0.825	0.805	0.898	0.908	0.859	0.814	0.909	0.897
w/ Contrast Pre-train	0.917	0.913	0.912	0.888	0.851	0.831	0.860	0.872	0.835	0.815	0.908	0.895
w/ Sharpness Pre-train	0.920	0.916	0.917	0.891	0.863	0.833	0.880	0.887	0.833	0.829	0.913	0.899
w/ Noisiness Pre-train	0.919	0.915	0.912	0.891	0.854	0.831	0.859	0.863	0.847	0.812	0.910	0.897
w/ ALL Pre-train	**0.922**	**0.918**	**0.923**	**0.901**	**0.873**	**0.855**	**0.934**	**0.930**	**0.861**	**0.846**	**0.919**	**0.906**

TABLE VIII

Performance of our SARQUE model using different feature fusing blocks on target databases: SPAQ [27], KonIQ-10k [26], LIVEW [48], CID2013 [50], and RBID [49].

Metric	SPAQ [27]		KonIQ-10k [26]		LIVEW [48]		CID2013 [50]		RBID [49]	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Concatenation [60]	0.919	0.915	0.918	0.898	0.852	0.824	0.870	0.875	0.845	0.833
Point-wise Addition [17]	0.918	0.915	0.919	0.897	0.854	0.823	0.866	0.863	0.853	0.835
Self-attention Fusion [61]	0.919	0.915	0.918	0.897	0.859	0.830	0.884	0.878	0.832	0.807
GCN [30]	**0.922**	**0.918**	**0.923**	**0.901**	**0.873**	**0.855**	**0.934**	**0.930**	**0.861**	**0.846**

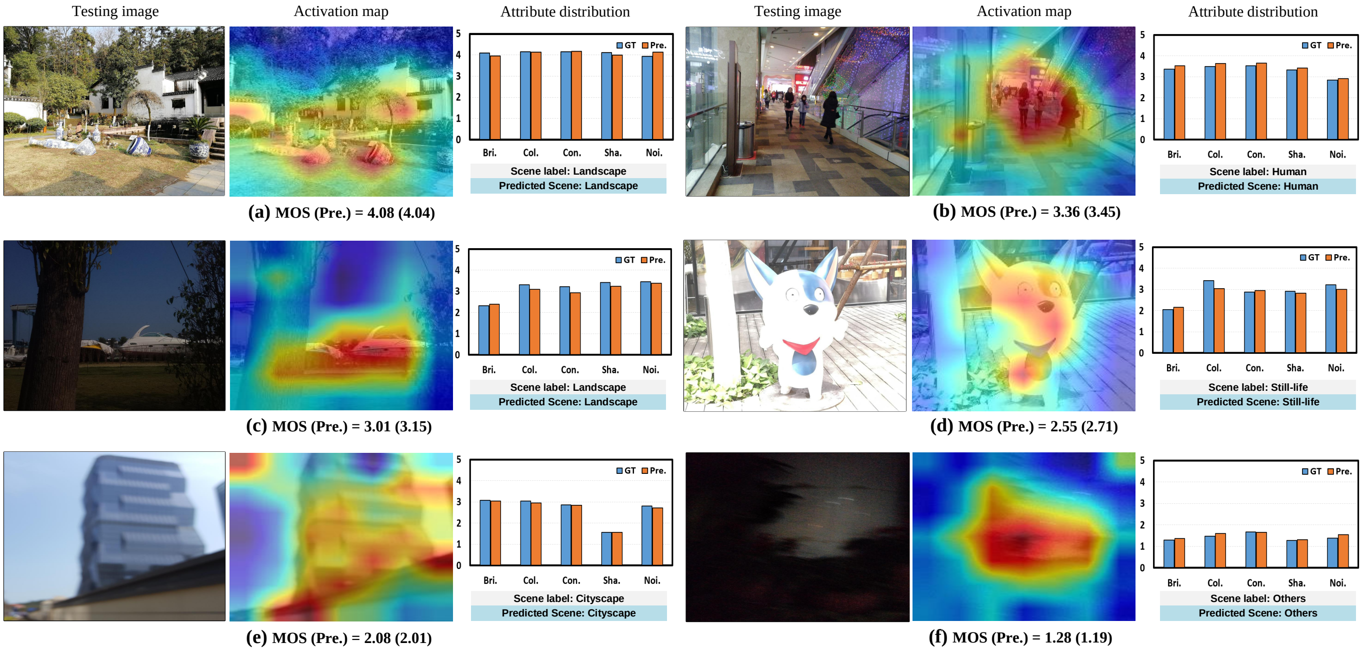
is more significant on three small databases including LIVEW, CID2013 and RBID. The possible reason is that these databases contain fewer images and scenes, which makes it difficult for the model to learn a comprehensive representation, so adding scene reasoning can bring more performance gain. This demonstrates the necessity of integrating the features from attributes and scene for reasoning the image quality score.

F. Effectiveness of pre-trained models

In this work, we first employ the SPAQ database to pre-train the semantic attribute learning and scene category prediction modules. Then, target databases are used to fine-tune all parameters for quality prediction. To explore the effectiveness of the pre-trained weights of semantic attribute learning and scene category prediction modules, we further conduct ablation experiments. First, we train and test SARQUE on five IQA datasets without loading any pre-trained weights for semantic attribute learning and scene category prediction modules (denoted as w/o Pre-train). Second, we load ImageNet [29] pretrained weights to semantic attribute learning and scene category prediction modules, and train and test on target database (denoted as w/ ImageNet Pre-train). Third, we load ImageNet pre-trained weights for the semantic attribute learning module and the proposed pre-trained weights for the scene category prediction module to test the performance of SARQUE (denoted as w/o Attributes Pre-train). Finally, we load the proposed pre-trained weights for semantic attribute learning and scene category prediction, and perform the same training and testing strategies (denoted as w/ Attribute+Scene Pre-train). The experimental results are listed in Table VI.

From the table, it can be observed that the proposed SARQUE loading attribute+scene pre-train weights overwhelmingly surpasses ImageNet pre-trained weights with significant margins for both PLCC and SRCC evaluations on all databases. In addition, SARQUE with ImageNet pretrained weights performs better than that no pre-trained weights, especially on small databases including LIVEW, CID2013, and RBID. These results further demonstrate that effective pre-training strategies are very important for designing deep learning-based IQA models.

As mentioned before, the semantic attributes are highly related to image quality. Therefore, during the pre-training process of semantic attribute learning, five quality-aware semantic attributes are used as the targets for model learning, including brightness, colorfulness, contrast, noisiness and sharpness. To



(a) **MOS (Pre.) = 4.08 (4.04)**

(b) **MOS (Pre.) = 3.36 (3.45)**

(c) **MOS (Pre.) = 3.01 (3.15)**

(d) **MOS (Pre.) = 2.55 (2.71)**

(e) **MOS (Pre.) = 2.08 (2.01)**

(f) **MOS (Pre.) = 1.28 (1.19)**

Fig. 6. Illustration of the six testing images, the corresponding activation maps via the commonly used CAM [62] method, and the corresponding predicted results of the proposed SARQUE model. Bri.: Brightness; Col.: Colorfulness; Con.: Contrast; Sha.: Sharpness; Noi.: Noisiness; GT: Ground Truth; Pre.: Prediction score.

TABLE IX
PERFORMANCE OF THE PROPOSED MODEL PER-TRAINED ON SPAQ DATABASE AND THEN FINE-TUNING (ONLY USING 20% IMAGES) ON OTHER DATABASES: KONIQ-10K [26], LIVEW [48], CID2013 [50], AND RBID [49].

Model	KonIQ-10k [26]		LIVEW [48]		CID2013 [50]		RBID [49]	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
fine-tune the pre-trained model using four FC layers	0.835	0.805	0.838	0.821	0.701	0.606	0.759	0.762
fine-tune the pre-trained model using GCN	**0.893**	**0.872**	**0.864**	**0.837**	**0.784**	**0.740**	**0.823**	**0.818**

further investigate the respective importance of the semantic attributes for quality prediction, we respectively pre-train a single attribute, and then fine-tune the entire network for the quality assessment. Table VII summarizes the experimental results. In general, pre-training for the sharpness task is most important for quality prediction. Specifically, only per-training the colorfulness task achieves the best performance on KonIQ-10k, CID2013 and RBID databases, and only per-training the sharpness task delivers the best performance on SPAQ and LIVEW databases. In addition, when all semantic attributes are combined to per-train the proposed model, the performance further improves significantly.

G. *Effectiveness of GCN*

To mine the interaction of semantic attributes and scene category in the IQA task, we propose a GCN-based semantic reasoning module. From a different point of view, GCN in our work is like a feature fusing module to integrate features of semantic attributes and scene category. To investigate the effectiveness of GCN, we further conduct comparison experiments by replacing GCN with different feature fusing strategies. Specifically, we compare three feature fusion strategies including Concatenation [60], Point-wise Addition [17] and Self-attention Fusion [61]. The results are summarized in Table VIII, where the best results are shown boldfaced. From Table VIII, we can observe that the proposed model equipped with GCN outperforms other feature fusion strategies on all databases, which further proves that using GCN for semantic reasoning can obtain better performance.

Furthermore, with the help of GCN, the proposed SARQUE model has achieved the best generalization performance in cross-database experiments without fine-tuning (in Table IV). In practice, a good BIQA model is also expected to have the capability of quickly adapting to a new BIQA task by performing simple fine-tuning using a small number of training samples. To explore the quick learning ability of GCN in the proposed model, we first train SARQUE on the SPAQ database. Then, we use only 20% of the data in a target database to fine-tune GCN or FC layers, and the remaining 80% images are used to test the performance of our SARQUE model. Moreover, for comparison, we use a commonly used MLP model to replace the GCN model and repeat this experiment, where the MLP model includes four FC layers with 256, 64, 8 and 1 nodes, respectively. The experimental results are listed in Table IX, where the best results are shown boldfaced. It can be seen from Table IX that GCN is advantageous over FC pooling by a sizable margin for all databases. In addition, compared with the cross-database experimental results without fine-tuning (in Table IV), the performance further improves significantly. A possible reason is that the GCN model takes into account the connection relationship between nodes and surrounding neighbor nodes in the process of convolution as shown in Equ. (9), while MLP does not. This confirms the advantage of the proposed model, which uses GCN to mine the inherent relationship between the semantic attributes and scene category.

H. Visual Analysis

To intuitively demonstrate the explainability of the proposed SARQUE model, we introduce a visual experiment on six testing images with different quality scores. Fig. 6 shows the testing images and the corresponding activation maps via the commonly used CAM [62] method based on the proposed semantic attribute learning module, as well as the corresponding predicted results of the proposed SARQUE model, in terms of quality score regression, semantic attribute prediction, and scene category recognition. It is worth noting that the ground truth of semantic attributes represents the human perception in images, not the absolute intensity in the physical meaning. For example, both overexposure and underexposure will bring bad visual experience, resulting in low brightness scores.

From Fig. 6, we have the following observations. 1) With the decreasing MOS values, the predicted quality scores also decrease accordingly. In addition, the predicted quality scores are very close to the ground truth. 2) The appearances of all activation maps are consistent with the position of attention when people judge image quality. 3) The proposed SARQUE model can effectively predict the sematic attributes of images, which are very close to the ground truth values. 4) The predicted semantic attributes can explain the dominant reason of image distortions. Specifically, all attribute values of Fig. 6(a) are very high, therefore it has a high quality score. For Fig. 6(b), the dominant reason that impacts the quality perception is noise, and the predicted noisiness value is also the lowest. The dominant distortions in Figs. 6(c) and 6(d) are underexposure and overexposure, respectively, and the predicted brightness values are both the lowest. In Fig. 6(e), the predicted sharpness value is the lowest, which is consistent with the human perception. For Fig. 6(f), due to the low attribute values, its quality score is also very low. From these visual results, we know that SARQUE can also provide reasonable explanations (in terms of semantic attributes) on why a particular quality score is predicted, which is highly desired in real-world applications.

V. Conclusion

In this paper, we have presented an explainable and generalizable BIQA model via semantic attribute reasoning, dubbed SARQUE. Different from the existing BIQA models, SARQUE can characterize the sematic attributes and scene category that determine image quality. The proposed two-stream network with multiple branches has demonstrated its effectiveness in learning sematic attributes and scene category. Further, by mining the inherent relationship between the semantic attributes and scene category, the proposed semantic reasoning module can accurately predict the image quality score. Extensive experiments conducted on five in-the-wild image quality databases have demonstrated that SARQUE is superior to the state-of-the-art BIQA models in terms of both evaluation accuracy and generalization ability. In addition, visual analysis have shown that SARQUE can provide the reasons for the degradation of image quality, which makes our model explainable.

References

[1] D. Li, T. Jiang, W. Lin, and M. Jiang, “Which has better visual quality: The clear blue sky or a blurry animal?” *IEEE Trans. Multimedia*, vol. 21, no. 5, pp. 1221–1234, May 2019.

[2] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, “Image quality assessment: Unifying structure and texture similarity,” *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, Early Access 2020.

[3] G. Zhai, Y. Zhu, and X. Min, “Comparative perceptual assessment of visual signals using free energy features,” *IEEE Trans. Multimedia*, vol. 23, pp. 3700–3713, Oct. 2021.

[4] J. Wu, J. Ma, F. Liang, W. Dong, G. Shi, and W. Lin, “End-to-end blind image quality prediction with cascaded deep neural network,” *IEEE Trans. Image Process.*, vol. 29, pp. 7414–7426, June 2020.

[5] D. Liu, P. An, R. Ma, W. Zhan, X. Huang, and A. A. Yahya, “Content-based light field image compression method with gaussian process regression,” *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 846–859, Aug. 2020.

[6] Z. Jin, M. Z. Iqbal, D. Bobkov, W. Zou, X. Li, and E. Steinbach, “A flexible deep cnn framework for image restoration,” *IEEE Trans. Multimedia*, vol. 22, no. 4, pp. 1055–1068, Aug. 2020.

[7] T. Dutta, A. Singh, and S. Biswas, “Styleguide: Zero-shot sketch-based image retrieval using style-guided image generation,” *IEEE Trans. Multimedia*, vol. 23, pp. 2833–2842, Aug. 2021.

[8] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, “Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment,” *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130–141, Nov. 2017.

[9] Z. Shao, J. Han, D. Marnerides, and K. Debattista, “Region-object relation-aware dense captioning via transformer,” *IEEE Trans. Neural Netw. Learn. Syst.*, pp. 1–12, Mar. 2022.

[10] Y. Liu, D. Zhang, Q. Zhang, and J. Han, “Part-object relational visual saliency,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 7, pp. 3688–3704, Jul. 2022.

[11] B. Yan, B. Bare, C. Ma, K. Li, and W. Tan, “Deep objective quality assessment driven single image super-resolution,” *IEEE Trans. Multimedia*, vol. 21, no. 11, pp. 2957–2971, May 2019.

[12] L. Li, W. Xia, W. Lin, Y. Fang, and S. Wang, “No-reference and robust image sharpness evaluation based on multiscale spatial and spectral features,” *IEEE Trans. Multimedia*, vol. 19, no. 5, pp. 1030–1040, Dec. 2017.

[13] L. Li, D. Wu, J. Wu, H. Li, W. Lin, and A. C. Kot, “Image sharpness assessment by sparse representation,” *IEEE Trans. Multimedia*, vol. 18, no. 6, pp. 1085–1097, Jun. 2016.

[14] A. Mittal, A. K. Moorthy, and A. C. Bovik, “No-reference image quality assessment in the spatial domain,” *IEEE Trans. Image Process.*, vol. 21, no. 12, pp. 4695–4708, Dec. 2012.

[15] Z. Wang, A. Bovik, H. Sheikh, and E. Simoncelli, “Image quality assessment: from error visibility to structural similarity,” *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[16] P. V. Vu and D. M. Chandler, “A fast wavelet-based algorithm for global and local image sharpness estimation,” *IEEE Signal Process. Lett.*, vol. 19, no. 7, pp. 423–426, May 2012.

[17] K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2016, pp. 770–778.

[18] Y. Huang, L. Li, H. Zhu, and B. Hu, “Blind quality index of depth images based on structural statistics for view synthesis,” *IEEE Signal Process. Lett.*, vol. 27, pp. 685–689, Apr. 2020.

[19] Y. Zhou, L. Li, S. Wang, J. Wu, Y. Fang, and X. Gao, “No-reference quality assessment for view synthesis using dog-based edge statistics and texture naturalness,” *IEEE Trans. Image Process.*, vol. 28, no. 9, pp. 4566–4579, Sep. 2019.

[20] L. Liu, Y. Hua, Q. Zhao, H. Huang, and A. C. Bovik, "Blind image quality assessment by relative gradient statistics and adaboosting neural network," *Signal Process. Image Commun.*, vol. 40, pp. 1–15, Jan. 2016.

[21] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2020, pp. 3664–3673.

[22] Y. Miao, Z. Lin, X. Ma, G. Ding, and J. Han, "Learning transformation-invariant local descriptors with low-coupling binary codes," *IEEE Trans. Image Process.*, vol. 30, pp. 7554–7566, Aug. 2021.

[23] H. Sheikh, M. Sabir, and A. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3440–3451, Oct. 2006.

[24] N. Ponomarenko, O. Ieremeiev, V. Lukin, K. Egiazarian, L. Jin, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C. C. J. Kuo, "Color image database TID2013: Peculiarities and preliminary results," in *Proc. Eur. Workshop Visual Inf. Process.*, Oct. 2013, pp. 106–111.

[25] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *J. Electron. Imaging*, vol. 19, no. 1, p. 011006, Jan. 2010.

[26] 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, Jan. 2020.

[27] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, June 2020, pp. 3674–3683.

[28] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Aug. 2020, pp. 14 131–14 140.

[29] I. S. A. Krizhevsky and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proc. Neural Inf. Process. Syst.*, Dec. 2012, pp. 1097–1105.

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

[31] 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.

[32] 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, Mar. 2013.

[33] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Trans. Image Process.*, vol. 24, no. 8, pp. 2579–2591, Aug. 2015.

[34] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, July 2012, pp. 1098–1105.

[35] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4444–4457, Sept. 2016.

[36] 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.*, June 2014, pp. 1733–1740.

[37] W. Hou, X. Gao, D. Tao, and X. Li, "Blind image quality assessment via deep learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 26, no. 6, pp. 1275–1286, Aug. 2015.

[38] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

[39] 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 and Syst. Video Technol.*, vol. 30, no. 1, pp. 36–47, Dec. 2020.

[40] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "RankIQA: Learning from rankings for no-reference image quality assessment," in *Proc. IEEE Int. Conf. on Comput. Vis.*, Oct. 2017, pp. 1040–1049.

[41] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipIQ: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Trans. Image Process.*, vol. 26, no. 8, pp. 3951–3964, May 2017.

[42] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "Generalizable no-reference image quality assessment via deep meta-learning," *IEEE Trans. Circuits and Syst. Video Technol.*, pp. 1–1, Early Access 2021.

[43] L. Leveque, J. Yang, X. Yang, P. Guo, K. Dasalla, L. Li, Y. Wu, and H. Liu, "CUID: A new study of perceived image quality and its subjective assessment," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2020, pp. 116–120.

[44] Y. Huang, J. Chen, W. Ouyang, W. Wan, and Y. Xue, "Image captioning with end-to-end attribute detection and subsequent attributes prediction," *IEEE Trans. Image Process.*, vol. 29, pp. 4013–4026, Jan. 2020.

[45] K. Liang, H. Chang, B. Ma, S. Shan, and X. Chen, "Unifying visual attribute learning with object recognition in a multiplicative framework," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 7, pp. 1747–1760, June 2019.

[46] P. Gong, X. Wang, Y. Cheng, Z. J. Wang, and Q. Yu, "Zero-shot classification based on multitask mixed attribute relations and attribute-specific features," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 12, no. 1, pp. 73–83, Feb. 2020.

[47] Z. Chen, X. Wei, P. Wang, and Y. Guo, "Learning graph convolutional networks for multi-label recognition and applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, pp. 1–1, Mar. 2021.

[48] 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, Nov. 2016.

[49] A. Ciancio, A. L. N. T. Targino da Costa, E. A. B. da Silva, A. Said, R. Samadani, and P. Obrador, "No-reference blur assessment of digital pictures based on multifeature classifiers," *IEEE Trans. Image Process.*, vol. 20, no. 1, pp. 64–75, June 2011.

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

[51] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Trans. Multimedia*, vol. 17, no. 1, pp. 50–63, Nov. 2015.

[52] S. Bosse, D. Maniry, T. Wiegand, and W. Samek, "A deep neural network for image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, Sept. 2016, pp. 3773–3777.

[53] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 1, pp. 206–220, Dec. 2017.

[54] S. Bosse, D. Maniry, K. R. Mller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[55] S. V. R. Dendi, C. Dev, N. Kothari, and S. S. Channappayya, "Generating image distortion maps using convolutional autoencoders with application to no reference image quality assessment," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 89–93, Nov. 2019.

[56] J. Kim, A.-D. Nguyen, and S. Lee, "Deep CNN-based blind image quality predictor," *IEEE Trans. Neural. Netw. Learn. Syst.*, vol. 30, no. 1, pp. 11–24, June 2019.

[57] B. Yan, B. Bare, and W. Tan, "Naturalness-aware deep no-reference image quality assessment," *IEEE Trans. Multimedia*, vol. 21, no. 10, pp. 2603–2615, Oct. 2019.

[58] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L. Li, "YFCC100M: The new data in multimedia research," *Commun. ACM*, vol. 59, no. 2, pp. 64–73, Jan. 2016.

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

[60] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput.-Assisted Intervention*, Oct. 2015, pp. 234–241.

[61] L. Li, T. Song, J. Wu, W. Dong, J. Qian, and G. Shi, "Blind image quality index for authentic distortions with local and global deep feature aggregation," *IEEE Trans. Circuits and Syst. Video Technol.*, pp. 1–1, Early Access 2021.

[62] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 2921–2929.

Yipo Huang (Student Member, IEEE) received the B.S. degree from Zhengzhou University, Zhengzhou, China, in 2017, and the M.S. degree from the China University of Mining and Technology, Xuzhou, China, in 2021. He is currently pursuing the Ph.D. degree with the School of Artificial Intelligence, Xidian University, Xi'an, China. His research interests include multimedia quality assessment, computational aesthetics and perceptual image processing.

Leida Li (Member, IEEE) received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2004 and 2009, respectively. In 2008, he was a Research Assistant with the Department of Electronic Engineering, National Kaohsiung University of Science and Technology, Kaohsiung, Taiwan. From 2014 to 2015, he was a Visiting Research Fellow with the Rapid-Rich Object Search Laboratory, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he was a Senior Research Fellow from 2016 to 2017. He is currently a Professor with the Guangzhou Institute of Technology, Xidian University, China. His research interests include multimedia quality assessment, affective computing, information hiding, and image forensics. He has served as an SPC for IJCAI 2019-2020, the Session Chair for ICMR in 2019 and PCM in 2015, and the TPC for AAAI in 2019, ACM MM 2019-2020, ACM MM-Asia in 2019, ACII in 2019, and PCM in 2016. He is currently an Associate Editor of the *Journal of Visual Communication and Image Representation* and the *EURASIP Journal on Image and Video Processing*.

Yuzhe Yang (Member, IEEE) received the B.S. degree from the China University of Mining and Technology, Xuzhou, China, in 2019. He received the M.S. degree from University of Southampton, Southampton, U.K., in 2020. Currently, he is a computer vision algorithm engineer at OPPO Research Institute. His research interests include multimedia affective computing, multimedia quality assessment and representation learning.

Yaqian Li received the B.S. degree from Lanzhou University, and the M.S. degree from Harbin Institute of Technology, China, in 2011 and 2013, receptively. He is currently the Technical Lead of visual search and understanding with OPPO Research Institute. His professional interests lie in the broad area of visual recognition, image retrieval, object detection, image quality assessment, and multimodality learning.

Yandong Guo received the B.S. and M.S. degrees in ECE from Beijing University of Posts and Telecommunications, China, in 2005 and 2008, receptively, and the Ph.D. degree in ECE from Purdue University at West Lafayette in 2013, under the supervision of Prof. Bouman and Prof. Allebach.

He is currently the Chief Scientist of Intelligent Perception with OPPO and chair the AI strategic planning for OPPO. He also holds an adjunct professor position at the Beijing University of Posts and Telecommunications. Before he joined OPPO in 2020, he was the Chief Scientist with XPeng Motors, China, and previously a researcher with Microsoft Research, Redmond, WA, USA. His professional interests lie in the broad area of computer vision, imaging systems, human behavior understanding and biometric, and autonomous driving.