Knowledge-Guided Blind Image Quality Assessment with Few Training Samples

Tianshu Song, Leida Li, Jinjian Wu, Yuzhe Yang, Yaqian Li, Yandong Guo, and Guangming Shi, *Fellow, IEEE*

***Abstract*—Blind image quality assessment (BIQA) for in-the-wild images has achieved great progress by training advanced deep neural networks. However, the current BIQA models are suffering the generalization challenge, meaning that a well-trained BIQA model is still very limited in evaluating images with different distributions. Deep BIQA models are data-intensive, but the annotation of image quality labels is extremely expensive. To design a generalizable BIQA model with few training samples is highly desired. Motivated by the above fact, this paper presents a knowledge-guided BIQA (KG-IQA) framework by integrating domain knowledge from the human visual system (HVS) and natural scene statistics (NSS). Specifically, the quality-aware HVS and NSS features are first extracted as prior knowledge. Then, we embed the two types of knowledge into the conventional deep neural network by learning to predict the HVS and NSS features, producing the knowledge-enhanced quality features, based on which the final image quality score is obtained. We conduct extensive experiments and comparisons on five authentically distorted IQA datasets. The experimental results demonstrate that the introduction of knowledge greatly reduces the requirement on the amount of training images, and the proposed KG-IQA model achieves superior performance in terms of both prediction accuracy and generalization ability.**

***Index Terms*—image quality assessment, knowledge embedding, human visual system, natural scene statistics, generalization.**

I. Introduction

BLIND image quality assessment (BIQA) has been a popular research field [1]–[8], which plays a vital role in image capture [9], [10], display [11], [12], and enhancement [13], *etc*. Due to the complexity of image distortions, designing BIQA models to meet the real-world applications is very challenging. Recent BIQA metrics have paid more attention to authentic distortions, and great progress has been achieved by leveraging advanced deep neural networks (DNNs).

BIQA metrics need to handle both challenges of diversified distortion types and varying visual contents, which are extremely difficult for traditional handcrafted feature-based models, *e.g.* BRISQUE [14], NFERM [15], and CORNIA [16]. Recent BIQA metrics adopt DNNs to solve the above challenges. DNNs follow the end-to-end learning paradigm, which have to be trained with a great number of labeled images. However, the amount of annotated training samples in BIQA is relatively small, so the existing deep BIQA models are subject to high risk of overfitting. This has led to the generalization challenge in BIQA, which has not drawn sufficient attention. When dealing with a new BIQA task, we typically have to label an extensive number of images. Each image needs to be annotated by many individuals (up to 120 annotators per image in [17]) to obtain the desired quality label, which is very expensive. An ideal BIQA model is expected to be trained with as few samples as possible, so that the model can be easily deployed for real-life applications at low cost. In recent years, knowledge-driven models have been widely studied in the conventional computer vision tasks, which consistently improves model performance by embedding domain knowledge [18]–[21]. However, the role of quality-aware prior knowledge in BIQA metrics has rarely been investigated so far.

Motivated by the above facts, this paper presents a new knowledge-guided deep BIQA framework, dubbed KG-IQA, with the objective to both reduce the amount of training samples and improve the generalization ability. Specifically, we introduce two types of domain knowledge into BIQA. First, the purpose of BIQA is to predict a quality score as close as possible to the human visual system (HVS). Thus, introducing the HVS knowledge into BIQA model is necessary. Second, the natural scene statistics (NSS) property represents the profound understanding about natural images and distortions from domain experts, and could also be introduced as the prior knowledge. In this paper, we embed both HVS and NSS knowledge into the proposed BIQA model. To integrate HVS knowledge, we first process the input images with the HVS properties of just noticeable difference (JND) and/or contrast sensitivity characteristics (CSC), producing the HVS-enriched images. Then, we extract features as knowledge from the HVS-enriched images through a Siamese-network. To integrate NSS knowledge, we extract NSS features as prior. After obtaining the domain knowledge, we further design a knowledge embedding module by learning to predict the knowledge-enriched features. By optimizing the feature prediction tasks and quality regression task simultaneously, the proposed KG-IQA model can be obtained.

The contributions of this paper are summarized as follows:

- We propose a new knowledge-guided BIQA framework for authentic distortions. KG-IQA features small sample training and better generalization ability, which are

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

Tianshu Song is with the School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China (e-mail: tianshusong@cumt.edu.cn).

Leida Li, Jinjian Wu, and Guangming Shi are with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mails: ldli@xidian.edu.cn, jinjian.wu@mail.xidian.edu.cn, gmshi@xidian.edu.cn).

Yuzhe Yang, Yaqian Li, and Yandong Guo are with the OPPO Research Institute, Shanghai 200032, China. (e-mails: ippllewis@gmail.com, liyaqian@oppo.com, yandong.guo@live.com).

achieved by integrating quality-aware prior knowledge from both the HVS and NSS.
- We propose a HVS knowledge representation method by pre-processing input images with HVS properties and then extracting HVS-enriched features from them. The knowledge is then embedded into the deep BIQA model through a feature prediction-based strategy.
- We have done extensive experiments and comparisons on five authentically distorted BIQA datasets. The results demonstrate that the proposed model significantly reduces the number of training samples, and achieves superior performance in terms of both prediction accuracy and generalization ability.

II. Related Work

A. Blind Image Quality Assessment

Traditional BIQA metrics typically train a regressor to obtain quality scores based on handcrafted features (*e.g.* BRISQUE [14], NFERM [15], HOSA [22], and CORNIA [16]). Handcrafted features have explicit meanings and have shown decent evaluating ability on synthetic distortions. However, the handcrafted features are usually extracted based on the still limited understanding of image distortions, which are not comprehensive and thus are relatively limited in evaluating the authentic distortions.

With the boom of deep learning, deep neural networks have been widely adopted for BIQA. Earlier BIQA metrics (*e.g.* BIECON [23], MEON [24], and WADIQaM-NR [25]) typically built relatively shallow networks. Due to the limited representation ability, these metrics still do not perform very well. Due to the small sample property, recent DNN-based BIQA metrics typically adopt the ImageNet pre-training and fine-tuning framework. For example, Zhang *et al.* [26] utilized two subnetworks to respectively evaluate synthetic and authentic distortions. The synthetic subnetwork was pre-trained with synthetically distorted images, and the authentic subnetwork was pre-trained on ImageNet. Su *et al.* [27] fused high-level semantic features and multi-scale content features obtained from different layers of the pre-trained ResNet-50 [28] to deal with the diversified contents in BIQA. These metrics have achieved great advances in evaluating authentic distortions.

In addition to the prediction accuracy, recent works also paid attention to the generalization ability. For example, Yan *et al.* [29] adopted a multi-task learning strategy to combine traditional handcrafted features with the neural network for improving the representation ability of BIQA. Zhu *et al.* [30] utilized meta-learning to extract meta-knowledge from different kinds of synthetic distortions, which benefits the subsequent fine-tuning on authentic distortions. Ma *et al.* [31], Liu *et al.* [32] and Zhang *et al.* [33] respectively adopted a remember-and-reuse network, a split-and-merge distillation strategy and a specially designed continual learning strategy to address the cross-distortion/scenario BIQA tasks. While these metrics have achieved notable success, the generalization ability of BIQA models is still an open challenge. Further, to the authors' best knowledge, the effectiveness of deep BIQA metrics with few training samples has not been studied before, which may restrict their real-world applications. Inspired by this, we propose the knowledge-guided BIQA framework to reduce the number of training samples and meantime improve the model generalization ability.

B. Human Visual System

The characteristics of the human visual system (HVS) are related to the perceptual aspects of brightness, contrast, texture, color, motion, *etc*, which consist of color and spatial processing mechanisms, perception properties of scale and depth, spatio-temporal response mechanisms, attention and eye movements mechanisms, visual masking characteristics, *etc.* [34]. HVS is the foundation of BIQA, and the objective of BIQA is to predict a quality score as close as possible to the HVS. Therefore, we introduce the knowledge of HVS into our model.

Among all HVS properties, the contrast sensitivity characteristics (CSC) and just noticeable difference (JND) are the two aspects mostly related to our metric. First, the CSC represents the dependence of the lowest recognizable contrast of periodic test-object on the spatial frequency [35]. Extensive studies have shown that the HVS is insensitive to high-frequency information of images [35], which has been the foundation of many image processing tasks such as image compression (JPEG/JPEG2000 standards are based on CSC). This property is also vital for IQA because distorted regions with different spatial frequencies may lead to completely different perceptual quality. Second, the JND reveals the limitation of human visual perception, which has also been widely adopted in many image processing fields (*e.g.* watermarking). The JND is significantly affected by the luminance, which can be deduced from the famous Weber's law that the ratio of the just noticeable illuminance change to the background illuminance is approximately constant [36]. Chou *et al.* [37] conducted perceptual experiments to measure the dependence of the JND threshold on background luminance. This dependence is important for IQA because distortions with different background luminance may lead to significantly different visual quality. Therefore, we introduce the HVS knowledge of CSC and/or JND to design our model.

C. Natural Scene Statistics

Natural scene statistics (NSS) play an important role in representing and processing natural images, which has drawn intensive attention and many important NSS properties have been discovered [38]. For example, after applying the Fourier transform on natural images, the spectrum power falls with the frequency. Scaling up and down the natural images, the marginal distributions of statistics remain unchanged. Inspired by these NSS properties, some distortion-aware NSS properties were successively proposed for BIQA. For example, Moorthy *et al.* [39] adopted a set of neighboring wavelet coefficients as a statistical description of distortions in the image. Mittal *et al.* [14] adopted statistical coefficients of the mean subtracted contrast normalized (MSCN) to describe distorted images, whose distributions vary as a function of distortion. Saad *et al.* [40] designed quality assessment features based on the

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3233244

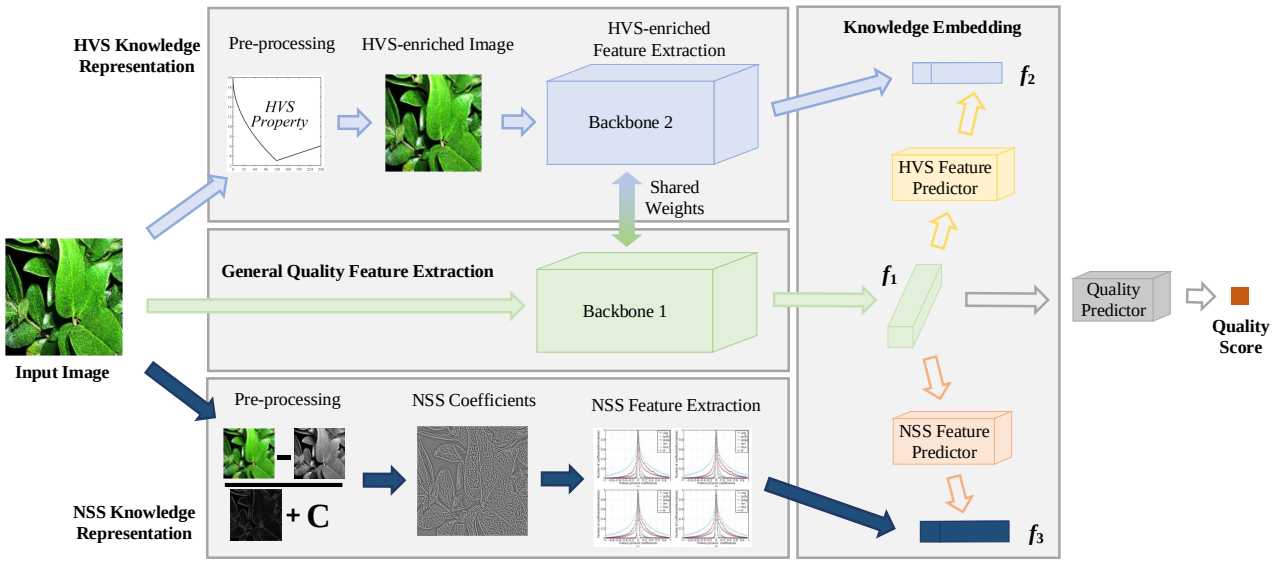


Fig. 1. The framework of the proposed Knowledge-Guided Image Quality Assessment (KG-IQA). The framework consists of four modules, including general quality feature extraction, HVS knowledge representation, NSS knowledge representation and knowledge embedding.

statistics of local discrete cosine transform (DCT) coefficients, which change with the image quality in a predictable manner. These NSS properties promoted the progress of BIQA to a great extent, and they represent profound understanding of natural images and distortions from domain experts. Therefore, we also introduce the NSS knowledge into our model.

III. Proposed Method

In this paper, to reduce the number of labeled training images and improve the generalization ability of BIQA models, we propose the Knowledge-Guided blind Image Quality Assessment (KG-IQA) framework, which is shown in Fig. 1. The KG-IQA consists of four modules, including general quality feature extraction, HVS knowledge representation, NSS knowledge representation and knowledge embedding. 1) The general quality feature extraction module extracts the general quality feature $\boldsymbol{f}_1$ from the input image through the backbone-1. 2) The HVS knowledge representation module first processes the input image with HVS properties and generates the HVS-enriched image. Then, the HVS-enriched image is sent to the backbone-2 (sharing weights with backbone-1) and generates the HVS-enriched feature $\boldsymbol{f}_2$. 3) The NSS knowledge representation module firstly performs pre-processing on the input image to obtain the NSS coefficients, and then extracts the NSS feature $\boldsymbol{f}_3$. 4) The knowledge embedding module embeds the knowledge into backbone-1 by training the general quality feature $\boldsymbol{f}_1$ to predict features $\boldsymbol{f}_2, \boldsymbol{f}_3$ through the HVS feature predictor and the NSS feature predictor. When embedding knowledge, the model simultaneously performs quality prediction through the quality predictor. Guided by the domain knowledge, the model is expected to achieve better prediction accuracy and generalization ability.

A. General Quality Feature Extraction

The general quality feature extraction module has the input of original images $\boldsymbol{I}$ and outputs the general quality feature $\boldsymbol{f}_1$ through the backbone-1:

$$\boldsymbol{f}_1 = B_1(\boldsymbol{I}), \tag{1}$$

where B_1 is the backbone-1. The general quality feature $\boldsymbol{f}_1$ will be utilized to predict both features $\boldsymbol{f}_2, \boldsymbol{f}_3$ and quality score after performing knowledge embedding.

B. HVS Knowledge Representation

The HVS knowledge representation module has the input of the original image and outputs the HVS-enriched feature. First, we generate the HVS-enriched image $\boldsymbol{I}_h$ from the input image $\boldsymbol{I}$ through the HVS property integration operation H:

$$\boldsymbol{I}_h = H(\boldsymbol{I}). \tag{2}$$

Then, we obtain the HVS-enriched feature $\boldsymbol{f}_2$:

$$\boldsymbol{f}_2 = B_2(\boldsymbol{I}_h), \tag{3}$$

where B_2 is the backbone-2. B_2 shares the same weights with the B_1, so that the differences between $\boldsymbol{f}_2$ and $\boldsymbol{f}_1$ originates from the differences between $\boldsymbol{I}$ and $\boldsymbol{I}_h$, which are the introduced knowledge.

Specifically, we propose two HVS property integration operations, including just noticeable difference (JND) and contrast sensitivity characteristics (CSC). In the proposed framework, both kinds of HVS knowledge are effective and can consistently improve the model performance. Therefore, in implementation, we can use any one of them in the proposed KG-IQA framework. As aforementioned, JND is significantly affected by luminance [37], which is measured as:

$$\boldsymbol{T}(x) = \begin{cases} 20 - 17\sqrt{\frac{\boldsymbol{B}(x)}{127}}, & if\ \boldsymbol{B}(x) < 127, \\ \frac{3}{128} + \frac{3}{128}\boldsymbol{B}(x), & else, \end{cases} \tag{4}$$

where $\boldsymbol{T}(x)$ is the threshold of JND at pixel x, and $\boldsymbol{B}(x)$ is the background luminance. When generating the JND-enriched image $\boldsymbol{I}_h$, we first calculate the threshold at each

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3233244

pixel of the image according to equation (4). Then, we introduce perturbations (smaller than the calculated JND) into the image to obtain the HVS-enriched image $\boldsymbol{I}_h$. The common strategies randomly inject the JND perturbations as noise [36], [41], which is impossible to be predicted for the knowledge embedding module. Different from noise injection operations, we introduce the JND perturbations at the most sensitive positions of the image to maximize the effectiveness of the introduced perturbations and the JND knowledge. Following the methods of generating adversarial examples [42], we first extract features from the input image $\boldsymbol{I}$ through backbone-2, and predict the quality scores S' through a quality predictor P_s:

$$\begin{cases} \boldsymbol{f}' = B_2(\boldsymbol{I}), \\ s' = P_s(\boldsymbol{f}'). \end{cases} \tag{5}$$

Next, we introduce large perturbations to the labeled scores y_i and obtain scores y_i', where $y_i' \neq y_i, i = 1, 2, ..., N$. Then, we calculate the loss between the predicted scores s' and y_i':

$$L = \frac{1}{n}\sum_{i=1}^{n} l'(s_i', y_i'), i = 1, 2, ..., N, \tag{6}$$

where l' is the loss function (we adopt mean square error function here). After that, we perform the back-proportion of the loss L and obtain the gradients $\boldsymbol{G}(x)$ at each pixel x of $\boldsymbol{I}$. Finally, we introduce the perturbations at positions where the gradients are larger than 0 (sensitive positions) and obtain the final HVS-enriched image $\boldsymbol{I}_h$:

$$\boldsymbol{I}_h(x) = \begin{cases} \boldsymbol{I}(x) + \boldsymbol{T}(x) \times sign(\boldsymbol{G}(x)), if\ \boldsymbol{G}(x) > 0, \\ \boldsymbol{I}(x), \quad else, \end{cases} \tag{7}$$

where $sign(\boldsymbol{G}(x)) = 1$, if $\boldsymbol{G}(x) > 0$. After obtaining $\boldsymbol{I}_h$, we extract the JND-enriched feature $\boldsymbol{f}_2$ through equation (3).

In addition to JND, we also introduce the CSC property that HVS is insensitive to high-frequency information. The CSC follows the contrast sensitivity function (CSF) [35]:

$$C(f_r) = (0.0499 + 0.2964 f_r) \times e^{(-0.114 f_r)^{1.1}}, \tag{8}$$

where f_r is the spatial frequency. To introduce the CSF to generate the HVS-enriched image $\boldsymbol{I}_h$, we utilize the well-known JPEG2000 [43] compression algorithm to process the input image, because JPEG2000 is based on the CSF, which reduces insensitive high-frequency information and maintains the visual quality at low compression ratios. After that, we extract the CSF-enriched feature $\boldsymbol{f}_2$ from $\boldsymbol{I}_h$ through equation (3). The JND and CSF integration process are illustrated in Fig. 2. In implementation, we control the strength of the preprocessing operations, so that the HVS-enriched images have nearly the same visual quality with input images because keeping same visual quality is necessary to make the knowledge valid.

C. NSS Knowledge Representation

The NSS knowledge representation module extracts the NSS feature $\boldsymbol{f}_3$ from the input image. Among popular statistical descriptions for image distortions, the mean subtracted contrast normalized (MSCN) coefficients are extracted in the spatial

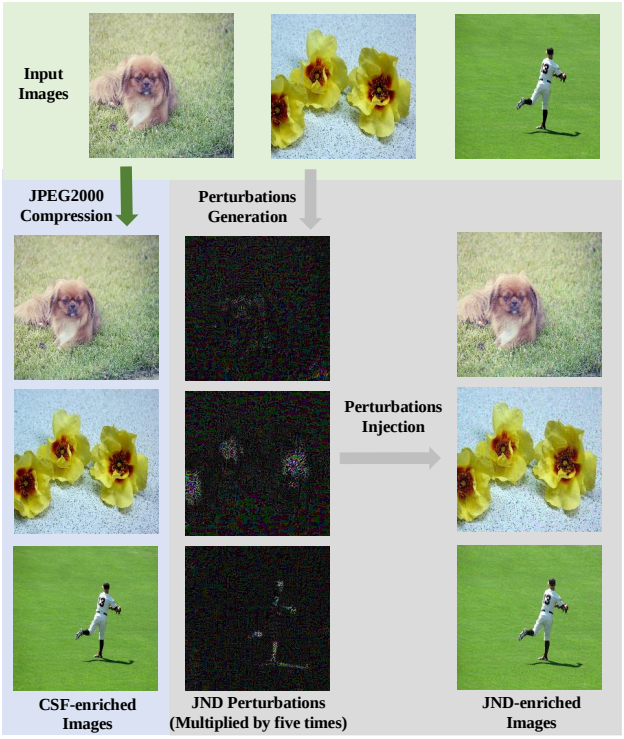


Fig. 2. HVS-enriched image generation process. The CSF-enriched images are obtained by compressing input images through JPEG2000 algorithms. The JND-enriched images are obtained by injecting JND perturbations into input images. We multiply the JND perturbation maps by five times for better display.

domain, which is highly efficient and can be calculated through a luminance transform operation [14]:

$$\boldsymbol{I}'(x) = \frac{\boldsymbol{I}(x) - \boldsymbol{\mu}(x)}{\boldsymbol{\sigma}(x) + \boldsymbol{C}}, \tag{9}$$

where $\boldsymbol{I}'(x)$ is the transformed luminance at pixel x, $\boldsymbol{\mu}(x), \boldsymbol{\sigma}(x)$ respectively denotes the local mean field and the local variation field, $\boldsymbol{C}$ is a constant to ensure numerical stability. Mittal *et al.* [14] defined the transformed luminances $\boldsymbol{I}'(x)$ as the MSCN coefficients, and observed that distributions of the MSCN coefficients vary as a function of distortions. Then, they adopted the generalized Gaussian distribution to fit the MSCN empirical distributions and obtain the final BIQA feature consisting of distribution parameters. This method is not only simple and effective but also represents a profound understanding of images distortions. Therefore, we adopt the NSS distribution parameters as the NSS feature $\boldsymbol{f}_3$.

D. Knowledge Embedding

To integrate the above domain knowledge into the proposed method, we further design a prediction-based knowledge embedding strategy. One may doubt that why not adopt the knowledge-enriched features $\boldsymbol{f}_2, \boldsymbol{f}_3$ as input and then fuse them with the general quality feature $\boldsymbol{f}_1$ to regress the final quality score. Though it seems more intuitive, simply adopting the knowledge-enriched features as input is inferior to predicting them [20], [21], [49]. First, instead of simply utilizing the knowledge-enriched features, the model can grasp the

TABLE I
DETAILED INFORMATION ABOUT FIVE AUTHENTIC BIQA DATASETS OF KONIQ-10K, SPAQ, LIVEW, CID2013 AND RBID.

Dataset	KonIQ-10k [17]	SPAQ [44]	LIVEW [45]	CID2013 [46]	RBID [47]
Number of Samples	10,073	11,125	1162	480	585
MOS Range	[1, 5]	[0, 100]	[0, 100]	[0, 5]	[0, 100]
Image Resolution	1024×768	1080×1440-6656×3744	500×500	1600×1200	480×640-2816×2112
Subject Environment	Crowdsourcing	Laboratory	Crowdsourcing	Laboratory	Laboratory
Number of Annotators	1459	N/A	8100	188	180
Number of Total Ratings	around 1.2 million	N/A	around 350,000	around 15,000	around 6,400
Other Labels	EXIF	EXIF, Attributes, Scene	N/A	Attributes	N/A
Image Source	Choosen from YFCC100m [48]	Taken with 66 mobile phones in the wild	Taken with 15 digital devices in the wild	Taken with 79 devices in eight scenes	Taken with 1 digital camera in the wild

knowledge by learning to predict them. For example, the HVS-enriched features contain HVS knowledge. When the BIQA model is able to precisely predict these features, the HVS knowledge within them has been embedded into the model. Similarly, if the model can precisely predict NSS features, the knowledge of obtaining NSS features has been successfully embedded into the model as well. The prior knowledge learned in the auxiliary task (feature prediction) benefits the main task (quality prediction) and leads to the better model performance, which has been proved by the previous multi-task learning research [49]. Second, if we adopt the knowledge-enriched features as input during training, we still need them at the test stage, which is time-consuming. In contrast, for the prediction strategy, knowledge-enriched features are used as labels during training, and they are not needed at the test stage, which is more effective for real-world applications. Therefore, we embed the knowledge through the feature prediction strategy.

The proposed knowledge embedding module consists of two predictors, *i.e.*, HVS knowledge predictor P_h and NSS knowledge predictor P_n. The module has inputs of the general quality feature $\boldsymbol{f}_1$, HVS-enriched feature $\boldsymbol{f}_2$, NSS knowledge-enriched feature $\boldsymbol{f}_3$, and the general quality feature $\boldsymbol{f}_1$ learns to predict the knowledge-enriched features $\boldsymbol{f}_2, \boldsymbol{f}_3$:

$$\begin{cases} \boldsymbol{f}_2' = P_h(\boldsymbol{f}_1), \\ \boldsymbol{f}_3' = P_n(\boldsymbol{f}_1). \end{cases} \tag{10}$$

The knowledge embedding module aims to embed the knowledge into the backbone, and the structure of predictors should be as simple as possible to ensure that not too much knowledge is embedded into the predictors.

E. Quality Score Regression

When embedding knowledge, the general quality feature $\boldsymbol{f}_1$ simultaneously learns to predict the quality score s_i through the quality predictor P_s:

$$s_i = P_s(\boldsymbol{f}_1), i = 1, 2, ..., N. \tag{11}$$

Then, we calculate the loss function and perform back-propagation:

Algorithm 1. Implementation details of the proposed KG-IQA.

Inputs: A batch of images $\boldsymbol{I}$, the target score of input images $y_i, i = 1, 2, ..., N$, where N is the batch size of training images.
Output: The total loss for back-propagation, denoted as L_t.

// Obtain original image feature $\boldsymbol{f}_1$ from $\boldsymbol{I}$:
$\boldsymbol{f}_1 \leftarrow Backbone1(\boldsymbol{I})$;
// Obtain HVS enriched input $\boldsymbol{I}_{hvs}$ from $\boldsymbol{I}$:
$\boldsymbol{I}_{hvs} \leftarrow HVS(\boldsymbol{I})$;
// Obtain HVS-enriched feature $\boldsymbol{f}_2$:
$\boldsymbol{f}_2 \leftarrow Backbone2(\boldsymbol{I}_{hvs})$;
// Obtain NSS knowledge-enriched feature $\boldsymbol{f}_3$:
$\boldsymbol{f}_3 \leftarrow NSS(\boldsymbol{I})$;
// Predict quality s through predictor P_s:
$s \leftarrow P_s(\boldsymbol{f}_1)$;
// Predict feature $\boldsymbol{f}_2'$ through predictor P_h:
$\boldsymbol{f}_2' \leftarrow P_h(\boldsymbol{f}_1)$;
// Predict feature $\boldsymbol{f}_3'$ through predictor P_n:
$\boldsymbol{f}_3' \leftarrow P_n(\boldsymbol{f}_1)$;
// Obtain the loss of quality score, denoted as L_1:
$L_1 \leftarrow \frac{1}{n}\sum_{i=1}^{n} l_1(s_i, y_i), i = 1, 2, ..., N$;
// Obtain the loss of $\boldsymbol{f}_2', \boldsymbol{f}_3'$, denoted as L_2, L_3:
$L_2 \leftarrow l_2(\boldsymbol{f}_2', \boldsymbol{f}_2)$, $L_3 \leftarrow l_2(\boldsymbol{f}_3', \boldsymbol{f}_3)$;
// Obtain the total loss L_t from L_1, L_2, L_3:
$L_t \leftarrow \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3$;

Return: L_t.

$$\begin{cases} L_1 = \frac{1}{n}\sum_{i=1}^{n} l_1(s_i, y_i), i = 1, 2, ..., N, \\ L_2 = l_2(\boldsymbol{f}_2', \boldsymbol{f}_2), \\ L_3 = l_2(\boldsymbol{f}_3', \boldsymbol{f}_3), \\ L_t = \lambda_1 L_1 + \lambda_2 L_2 + \lambda_3 L_3, \end{cases} \tag{12}$$

where y_i is the target quality score. l_1, l_2 denote the loss functions for quality regression and feature prediction, and L_1, L_2, L_3 are the obtained loss values. $\lambda_1, \lambda_2, \lambda_3$ are weight parameters, and L_t is the total loss for back-propagation. The implementation details of the proposed KG-IQA model are depicted in Algorithm 1.

After training, the model has grasped the knowledge by adopting the feature prediction strategy, we do not need the knowledge representation and embedding module during the test stage. In other words, the general quality feature $\boldsymbol{f}_1$ is directly sent into the quality predictor P_s and then generates the final quality score s_i through equation (11).

TABLE II
PLCC/SRCC RESULTS OF DIFFERENT AMOUNTS OF TRAINING IMAGES ON THE KONIQ-10K DATASET. THE MODELS ARE TRAINED WITH 5%, 10%, AND 25% IMAGES, AND TESTED ON THE REST IMAGES. THE BEST RESULTS ARE HIGHLIGHTED BOLD.

Amount of Training Images		5%	10%	25%
PLCC	NFERM [15]	0.615	0.651	0.687
	BRISQUE [14]	0.594	0.627	0.666
	CORNIA [16]	0.721	0.743	0.765
	HOSA [22]	0.730	0.751	0.777
	WaDIQaM-NR [25]	0.678	0.723	0.789
	DBCNN [26]	0.829	0.843	0.868
	MetaIQA [30]	0.796	0.821	0.861
	HyperNet [27]	0.800	0.842	0.883
	KG-IQA (NSS+JND)	0.848	**0.877**	0.897
	KG-IQA (NSS+CSF)	**0.850**	0.876	**0.901**
SRCC	NFERM [15]	0.588	0.622	0.656
	BRISQUE [14]	0.561	0.592	0.628
	CORNIA [16]	0.682	0.701	0.720
	HOSA [22]	0.685	0.708	0.737
	WaDIQaM-NR [25]	0.649	0.698	0.763
	DBCNN [26]	0.811	0.828	0.852
	MetaIQA [30]	0.761	0.788	0.830
	HyperNet [27]	0.768	0.814	0.859
	KG-IQA (NSS+JND)	0.816	**0.851**	0.874
	KG-IQA (NSS+CSF)	**0.825**	**0.851**	**0.877**

TABLE III
PLCC/SRCC RESULTS OF DIFFERENT AMOUNTS OF TRAINING IMAGES ON THE SPAQ DATASET. THE MODELS ARE TRAINED WITH 5%, 10%, AND 25% IMAGES, AND TESTED ON THE REST IMAGES. THE BEST RESULTS ARE HIGHLIGHTED BOLD.

Amount of Training Images		5%	10%	25%
PLCC	NFERM [15]	0.734	0.768	0.794
	BRISQUE [14]	0.737	0.762	0.790
	CORNIA [16]	0.812	0.827	0.843
	HOSA [22]	0.806	0.827	0.848
	WaDIQaM-NR [25]	0.809	0.831	0.861
	DBCNN [26]	0.873	0.885	0.898
	MetaIQA [30]	0.875	0.887	0.898
	HyperNet [27]	0.867	0.885	0.901
	KG-IQA (NSS+JND)	**0.890**	**0.900**	**0.911**
	KG-IQA (NSS+CSF)	0.883	0.895	**0.911**
SRCC	NFERM [15]	0.730	0.763	0.787
	BRISQUE [14]	0.733	0.756	0.783
	CORNIA [16]	0.805	0.820	0.836
	HOSA [22]	0.800	0.821	0.842
	WaDIQaM-NR [25]	0.806	0.827	0.857
	DBCNN [26]	0.874	0.885	0.900
	MetaIQA [30]	0.872	0.885	0.895
	HyperNet [27]	0.867	0.885	0.899
	KG-IQA (NSS+JND)	**0.885**	**0.895**	**0.908**
	KG-IQA (NSS+CSF)	0.877	0.891	0.907

IV. EXPERIMENTS

A. Evaluation Protocol

In our experiments, we adopt five authentically distorted BIQA datasets, including KonIQ-10k [17], LIVE in the Wild Image Quality Challenge (LIVEW) [45], RBID [47], Smartphone Photography Attribute and Quality (SPAQ) [44], and CID2013 [46]. Table I summarizes the detailed information of the five databases, including number of samples, MOS range, image resolution, image source, *etc.*

In implementation, we adopted different backbones to build the KG-IQA model, including VGG16 [50], ResNet18 [28], ResNet50 [28] and DeiT-small [51] (the default backbone). Both feature predictors and quality predictor consist of three fully connected (FC) layers with rectified linear unit (RELU) as the activation function. During training, both backbones and predictors share the same experimental settings. The loss functions for L_1 and L_2 are the mean square error (MSE). The weight parameter λ_1 is set to 1. The parameters λ_2, λ_3 respectively vary from $0.5-1.5$ and $0.5-2.5$ respectively according to different datasets and different backbones, which are determined by experiments. The optimizer we adopted is stochastic gradient descent (SGD) with initial learning rate of 0.01 and warm-up strategy (warm with 0.001 and 0.003 for 10 epochs respectively). When the loss does not decrease for 10 epochs, we decrease the learning rate by multiplying 0.3 until the learning rate is smaller than 1×10^{-5}. During training, we first resize the training images into size of 244×244, and then randomly crop a 224×224 input as the augmentation. At the test stage, we directly resize the image into 224×224.

During performance evaluation, we adopt popular evaluation criteria of Pearson Linear Correlation Coefficient (PLCC) and Spearman Rank-order Correlation Coefficient (SRCC). For the

TABLE IV
PLCC/SRCC RESULTS WITH 25% TRAINING IMAGES ON LIVEW, CID2013 AND RBID. THE MODELS ARE TRAINED WITH 25% IMAGES, AND TESTED ON THE REST IMAGES. THE BEST RESULTS ARE HIGHLIGHTED BOLD.

Evaluation Datasets		LIVEW	CID2013	RBID
PLCC	NFERM [15]	0.447	0.711	0.462
	BRISQUE [14]	0.477	0.709	0.492
	CORNIA [16]	0.635	0.744	0.647
	HOSA [22]	0.617	0.744	0.639
	WaDIQaM-NR [25]	0.538	0.732	0.507
	DBCNN [26]	0.721	0.761	0.745
	MetaIQA [30]	0.780	0.848	0.727
	HyperNet [27]	0.767	0.846	0.689
	KG-IQA (NSS+JND)	0.823	**0.859**	**0.781**
	KG-IQA (NSS+CSF)	**0.839**	0.858	0.764
SRCC	NFERM [15]	0.424	0.728	0.450
	BRISQUE [14]	0.456	0.724	0.489
	CORNIA [16]	0.587	0.732	0.632
	HOSA [22]	0.579	0.753	0.617
	WaDIQaM-NR [25]	0.512	0.717	0.491
	DBCNN [26]	0.754	0.696	0.758
	MetaIQA [30]	0.748	0.836	0.716
	HyperNet [27]	0.739	0.838	0.682
	KG-IQA (NSS+JND)	0.782	**0.847**	**0.761**
	KG-IQA (NSS+CSF)	**0.803**	0.845	0.744

predicted scores $\{s_1, s_2, ..., s_n\}$ and the labels $\{y_1, y_2, ..., y_n\}$, PLCC and SRCC are calculated by:

$$\mathrm{PLCC} = \frac{n\sum s_i y_i - \sum s_i \sum y_i}{\sqrt{n\sum s_i^2 - (\sum s_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}, \quad (13)$$

$$\mathrm{SRCC} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \quad (14)$$

where d_i is the difference of ranks of two sequences.

TABLE V
PLCC/SRCC RESULTS WITH 80% TRAINING IMAGES ON FIVE BIQA DATASETS. THE HVS KNOWLEDGE IN KG-IQA IS JND. THE BEST RESULTS ARE HIGHLIGHTED BOLD.

Dataset	KonIQ-10k [17]		SPAQ [44]		LIVEW [45]		CID2013 [46]		RBID [47]		Weighted Average	
Criteria	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
NFERM [15]	0.725	0.689	0.832	0.823	0.562	0.517	0.825	0.823	0.585	0.559	0.766	0.744
BRISQUE [14]	0.689	0.647	0.832	0.822	0.574	0.557	0.810	0.814	0.617	0.594	0.752	0.728
CORNIA [16]	0.773	0.738	0.867	0.859	0.692	0.655	0.822	0.803	0.712	0.695	0.813	0.792
HOSA [22]	0.791	0.761	0.873	0.866	0.703	0.667	0.835	0.833	0.716	0.684	0.825	0.806
NSSADNN [29]	/	/	/	/	0.813	0.745	0.825	0.748	/	/	0.817	0.746
MEON [24]	/	/	/	/	0.693	0.688	0.703	0.701	/	/	0.696	0.692
BIECON [23]	/	/	/	/	0.613	0.595	0.620	0.606	/	/	0.615	0.598
Zhang *et al.* 2021 [52]	/	0.847	/	/	/	0.835	/	/	/	0.827	/	0.845
CONTRIQUE [53]	0.906	0.894	0.919	0.914	0.857	0.845	/	/	/	/	0.910	0.901
UNIQUE [54]	0.901	**0.896**	/	/	**0.890**	0.854	/	/	**0.873**	**0.858**	0.899	0.890
HyperNet [27]	0.917	0.894	0.915	0.912	0.858	0.835	**0.922**	**0.913**	0.826	0.811	0.911	0.898
WaDIQaM-NR [25]	0.805	0.797	0.887	0.882	0.680	0.671	0.868	0.854	0.742	0.725	0.838	0.831
DBCNN [26]	0.884	0.875	0.915	0.911	0.869	0.851	0.871	0.863	0.859	0.845	0.897	0.890
MetaIQA [30]	0.887	0.850	0.871	0.870	0.835	0.802	0.784	0.766	0.777	0.746	0.872	0.853
KG-IQA	**0.918**	**0.896**	**0.922**	**0.919**	0.885	**0.862**	0.916	0.906	0.868	0.845	**0.917**	**0.904**

TABLE VI
PLCC/SRCC RESULTS OF CROSS-DATASET TEST. ALL MODELS ARE TRAINED ON 5% IMAGES OF KONIQ-10K AND DIRECTLY TESTED ON OTHER DATASETS. THE HVS KNOWLEDGE IN KG-IQA IS JND. THE BEST RESULTS ARE HIGHLIGHTED BOLD.

Dataset	SPAQ [44]		LIVEW [45]		CID2013 [46]		RBID [47]	
Criteria	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
NFERM [15]	0.622	0.629	0.471	0.468	0.463	0.346	0.418	0.483
BRISQUE [14]	0.556	0.563	0.456	0.439	0.542	0.559	0.482	0.509
CORNIA [16]	0.722	0.751	0.624	0.590	0.445	0.315	0.643	0.638
HOSA [22]	0.745	0.764	0.629	0.584	0.484	0.350	0.652	0.633
WaDIQaM-NR [25]	0.705	0.730	0.524	0.491	0.527	0.415	0.530	0.557
DBCNN [26]	0.822	0.815	0.713	0.692	0.734	0.650	0.714	0.721
MetaIQA [30]	0.781	0.795	0.709	0.687	0.707	0.664	0.674	0.677
HyperNet [27]	0.794	0.810	0.700	0.667	0.694	0.664	0.672	0.692
KG-IQA	**0.843**	**0.842**	**0.739**	**0.724**	**0.758**	**0.729**	**0.732**	**0.742**

B. Performance Comparison

The primary goal of our work is to improve the model performance with fewer training samples. Therefore, different from the common setting of training with 80% dataset images and testing on the rest 20% images, we train models using much smaller proportions (5%, 10% and 25%) of images from the whole databases, and test them on the rest images (accordingly 95%, 90% and 75%). We randomly split training and testing samples 10 times and the average results are reported. For fair comparison, we retrain the popular BIQA models following the same setting, and perform the same experiments. The following BIQA models are compared in our experiments, including the traditional handcrafted feature-based models of NFERM [15], BRISQUE [14], CORNIA [16], HOSA [22], and deep learning-based metrics of WaDIQaM-NR [25], DBCNN [26], MetaIQA [30], and HyperNet [27].

We first show the results of all metrics on the KonIQ-10k dataset in Table II. It is easily observed from Table II that the performances of all metrics significantly decrease when training with fewer images, especially when training with only 5%/10% labeled images. Guided by the knowledge of both HVS and NSS, the proposed KG-IQA consistently outperforms other metrics by a large margin. Among the compared metrics, both DBCNN and MetaIQA were pre-trained with many extra images with quality labels. Specifically, one subnetwork in DBCNN is pre-trained with a great number (more than 850 thousand) of synthetically distorted images with labels of distortion level and distortion type. MetaIQA is pre-trained on synthetically distorted BIQA dataset of Kadid-10k [57], which contains 25 types of distortions and 10125 images with quality labels. Though they were pre-trained with extensive quality labels, the proposed metric still achieves better performance by embedding the knowledge of HVS and NSS.

Then, we conduct the same experiments on the SPAQ dataset and the results are listed in Table III. The proposed KG-IQA also outperforms other metrics. By embedding knowledge, KG-IQA shows more obvious advantage at smaller ratios of 5% and 10%, which is consistent with results in Table II, indicating that embedding knowledge is more helpful when training samples are fewer.

To further demonstrate the effectiveness of the KG-IQA, we also train models on three relatively small datasets of LIVEW, CID2013 and RBID. It is known from Table I that the sales of the above three databases are much smaller than that of KonIQ-10k and SPAQ. Therefore, we only train models with

TABLE VII
PLCC/SRCC RESULTS OF CROSS-DATASET TEST. ALL MODELS ARE TRAINED ON 80% IMAGES OF KONIQ-10K AND DIRECTLY TESTED ON OTHER DATASETS. THE HVS KNOWLEDGE IN KG-IQA IS JND. THE BEST RESULTS ARE HIGHLIGHTED BOLD.

Dataset	SPAQ [44]		LIVEW [45]		CID2013 [46]		RBID [47]	
Criteria	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
NFERM [15]	0.688	0.711	0.548	0.540	0.715	0.680	0.520	0.530
BRISQUE [14]	0.650	0.682	0.575	0.554	0.555	0.533	0.581	0.597
CORNIA [16]	0.711	0.766	0.672	0.639	0.605	0.538	0.686	0.688
HOSA [22]	0.731	0.771	0.675	0.652	0.690	0.664	0.692	0.679
DeepRN (ResNet101) [55]	/	/	0.750	0.726	/	/	/	/
DeepBIQ (InceptionV2) [56]	/	/	0.821	0.804	/	/	/	/
ConCept512 [17]	/	/	0.848	0.825	/	/	/	/
UNIQUE [54]	/	/	/	0.786	/	/	/	0.783
HyperNet [27]	0.856	0.861	0.828	0.804	0.791	**0.755**	0.808	0.798
WaDIQaM-NR [25]	0.743	0.779	0.653	0.647	0.702	0.676	0.629	0.659
DBCNN [26]	0.851	0.850	0.764	0.729	0.781	0.736	0.777	0.784
MetaIQA [30]	0.834	0.851	0.806	0.783	0.764	0.710	0.780	0.781
KG-IQA	**0.862**	**0.871**	**0.837**	**0.805**	**0.794**	0.746	**0.823**	**0.818**

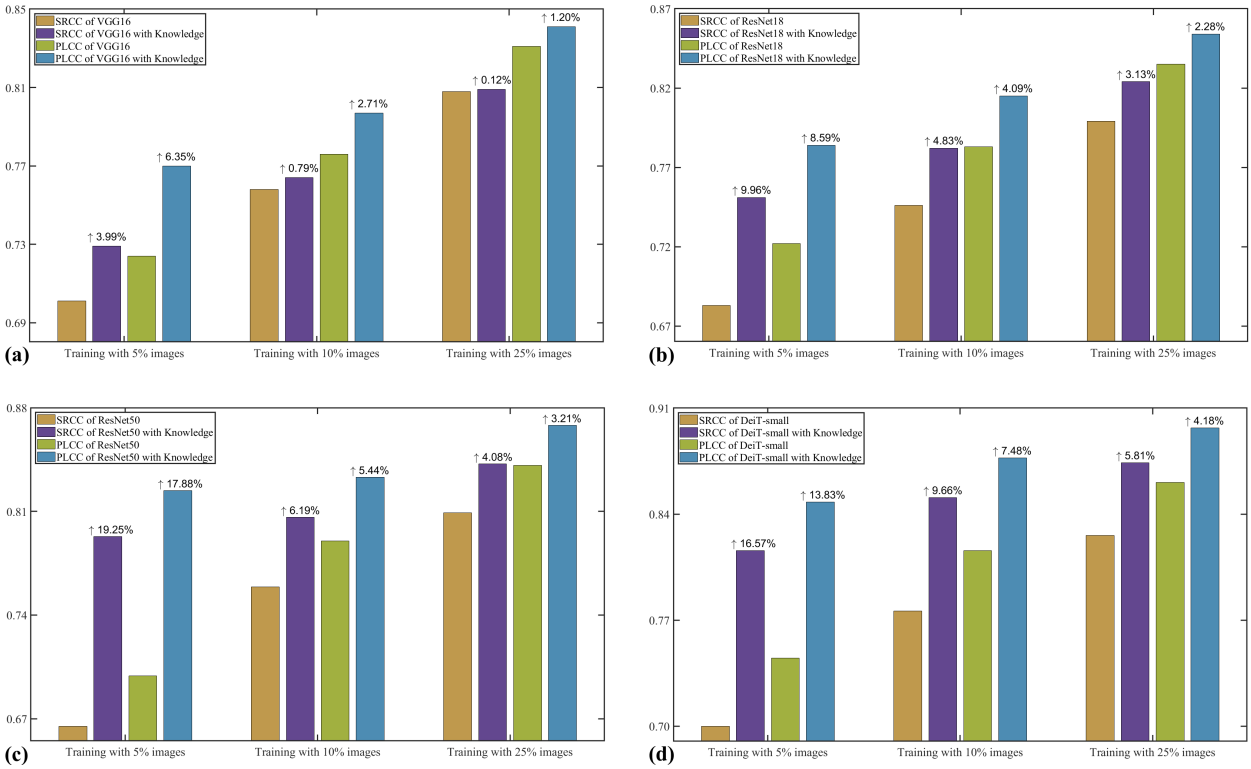


Fig. 3. Ablation results of the proposed KG-IQA with different backbones based on the KonIQ-10k dataset. (a), (b), (c), (d) respectively shows the results on backbones of VGG16, ResNet18, ResNet50, and DeiT-small. Each backbone is trained with/without knowledge on different ratios of 5%, 10% and 25% images on the KonIQ-10k dataset. The percentages above bars show the performance gain of PLCC/SRCC values after introducing knowledge.

25% labeled images on these three datasets, and the results are summarized in Table IV. It can be seen from Table IV that the proposed metric achieves the best performance, which also proves that embedding knowledge is an effective way to improve the evaluation ability for small-scale databases. We can also observe from Tables II-IV that embedding HVS knowledge of JND or NSS achieves similar results. Therefore, we choose the JND knowledge as the default HVS knowledge in the following experiments.

Besides training with few samples, we also train models with the conventional setting, where 80% database images are used for model training and the rest 20% images are used for test. The results are summarized in Table V, including the above eight retrained metrics and another six deep learning-based metrics of NSSADNN [29], MEON [24], BIECON [23], Zhang *et al.*2021 [52], CONTRIQUE [53], and UNIQUE [54] (results are obtained from published papers). From Table V, we can observe that the proposed KG-IQA achieves the best performance in most cases and the average performance of KG-IQA is also the best. Some compared metrics, *e.g.* HyperNet (fusing high-level semantic features and multi-scale content features to deal with the diversified contents), DBCNN

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3233244

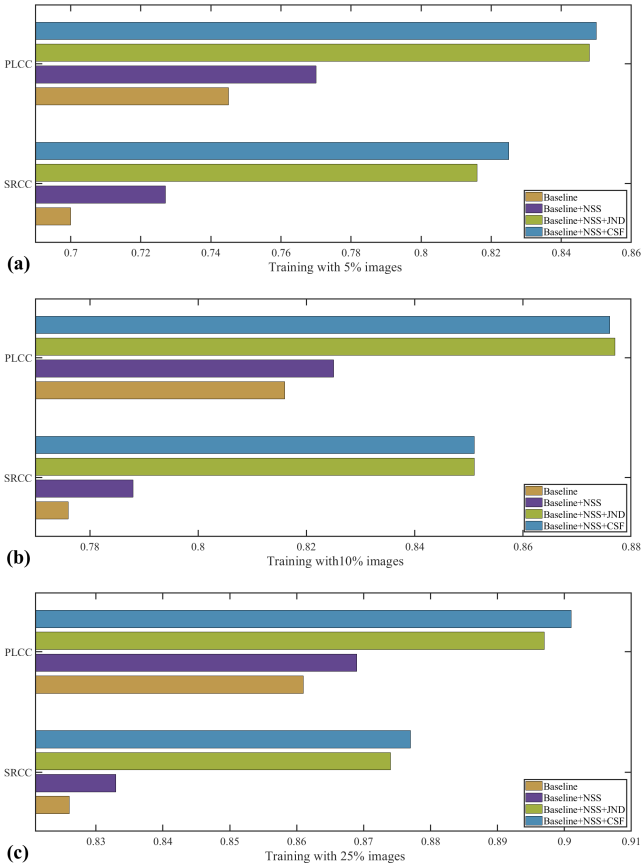


Fig. 4. Ablation results of each kind of knowledge. (a), (b), (c) respectively shows the results trained with 5%, 10%, 25% images on KonIQ-10k. The baseline models are trained without embedding knowledge.

(utilizing two subnetworks to respectively evaluate synthetic and authentic distortions), and UNIQUE (trained with a mixed dataset with more than 23k labeled images), also achieve very encouraging performance when training with 80% images.

C. Generalization Ability

As aforementioned, besides the prediction ability with fewer training samples, the generalization ability is another important factor of BIQA metrics. To investigate the generalization ability of the proposed metric, we train our model on one dataset and directly test it on other datasets. Considering its diversified image sources and relatively large database scale, we choose the KonIQ-10k as the training dataset and test it on the other four datasets. In this experiment, we use 5% of the images in KonIQ-10k to train the proposed model, and then it is used to test the performance of the other four databases directly. For comparison, we retrain the eight existing models under the same setting. Table VI summarizes the experimental results. It can be concluded from Table VI that by embedding knowledge of HVS and NSS, the generalization ability of the proposed KG-IQA outperforms previous metrics by a large margin. As aforementioned, one subnetwork in DBCNN is pre-trained with a great number (more than 850 thousand) of synthetically distorted images with labels of distortion level and distortion type. Benefiting from the extensive pre-training images, DBCNN also achieves very encouraging performance.

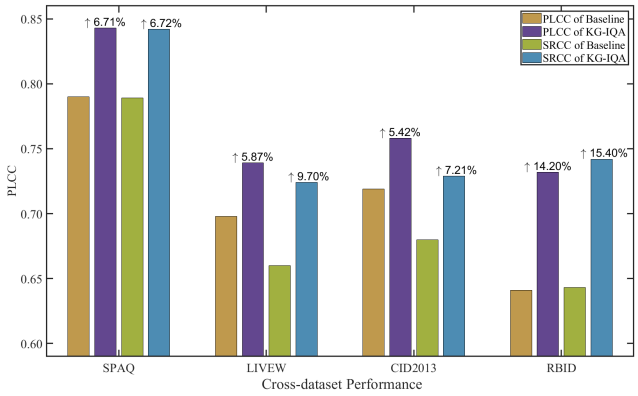


Fig. 5. Ablation results of generalization ability. The baseline model and KG-IQA are trained with 5% images on KonIQ-10k, and the baseline model is trained without knowledge.

We also train models with the common setting of 80% images on KonIQ-10k, and directly test them on other datasets. The results are listed in Table VII. From Table VII, we can observe that the proposed KG-IQA trained with 80% images can also outperform previous metrics in almost all cases. HyperNet, MetaIQA and DBCNN also achieves good generalization ability when training with 80% images.

D. Ablation Studies

To verify the universality of knowledge in the proposed KG-IQA model, we test the model performance with different backbone networks, including VGG16 [50], ResNet18 [28], ResNet50 [28] and DeiT-small [51]. The results trained on different numbers of images on KonIQ-10k with or without embedding knowledge are respectively shown in Fig. 3. It is observed from Fig. 3 that embedding HVS and NSS knowledge can significantly improve the evaluation ability with fewer training samples, which is agnostic to backbone networks. In other words, the HVS and NSS knowledge can be integrated to improve model performance with different backbones. The performance gain of models trained with 5% images is more significant than that of models trained with 10% and 25% images. This phenomenon is consistent with the above experiments in Tables II-III, indicating that knowledge plays a more important role when training with fewer samples. Consequently, we can conclude that embedding knowledge is an effective approach to reduce the training samples.

In this work, we propose two types of HVS knowledge, including JND and CSF. We also conduct an ablation study to show their respective contributions to the model performance. We first train models without knowledge as the baseline and then embed different kinds of knowledge into the baseline model. All results are shown in Fig. 4. We can observe from Fig. 4 that both NSS and HVS knowledge significantly improve the evaluation ability with fewer training samples. By embedding both HVS and NSS knowledge, the proposed framework achieves the best results.

In addition to the ablation experiments for intra-dataset tests, we also perform ablation study for the generalization ability. To be specific, we train models with or without knowledge

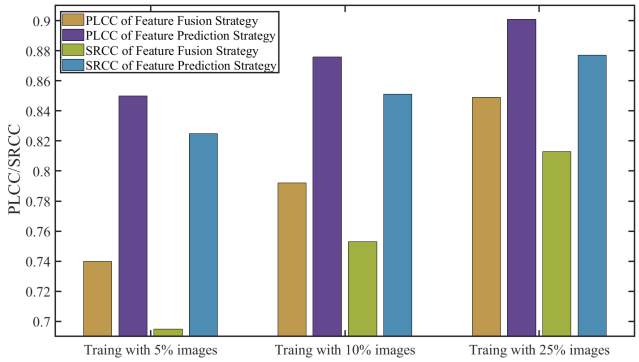


Fig. 6. Performance comparison of embedding knowledge through feature fusion and feature prediction. The models are trained with different number of images on KonIQ-10k.

on 5% images of KonIQ-10k, and directly test them on the other four datasets. The results are shown in Fig. 5. We can observe from Fig. 5 that embedding knowledge significantly improves the generalization ability of the model trained with fewer samples.

As aforementioned, we embed the knowledge by learning to predict the knowledge-enriched features. A more straightforward strategy is to directly utilize the knowledge-enriched features $\boldsymbol{f}_2, \boldsymbol{f}_3$ as input, and then fuse them with the general quality feature $\boldsymbol{f}_1$. The regressor can make quality prediction through the fused feature directly. Though this fusion strategy seems more intuitive, it has disadvantages. First, by adopting the knowledge-guided feature as input, the model only learns to combine the knowledge with general quality features instead of grasping the knowledge. Second, if we adopt the knowledge-enriched features as input, we also need them at the test stage, which is also time-consuming. To intuitively show the advantage of the proposed knowledge embedding strategy, we also conduct experiments by fusing the knowledge-enriched features $\boldsymbol{f}_2, \boldsymbol{f}_3$ with the original image feature $\boldsymbol{f}_1$ through concatenating operation with different amounts of training images. The ablation results are shown in Fig. 6. It is observed from Fig. 6 that compared with the feature fusion strategy, embedding knowledge through the prediction strategy is more effective.

V. Conclusions

In this paper, we have proposed a knowledge-guided BIQA framework to embed both HVS and NSS knowledge into deep neural network-based BIQA. We embed the domain knowledge by learning to predict the knowledge-enriched features. Experiments on five authentically distorted BIQA datasets show that the proposed metric achieves the best prediction accuracy and cross-dataset performance with much fewer training samples. When the amount of training samples becomes fewer, the knowledge plays a more vital role in improving both prediction accuracy and generalization ability, which indicates that embedding knowledge into BIQA model is an effective way to alleviate the dependence of DNNs on on annotated training samples.

Other kinds of HVS knowledge can be introduced in the proposed framework if the HVS-enriched images can be generated. For example, we can first generate HVS-enriched images by adopting some color-based HVS properties and then embed the knowledge into the neural network through the proposed knowledge embedding strategy. However, the proposed method has the limitation that some HVS properties, *e.g.* temporal response property of HVS, may be difficult to generate corresponding HVS-enriched images. Some other HVS-properties, *e.g.* attention mechanisms of HVS, the generated HVS-enriched images may be heat-map images, which is unsuitable for the backbone to extract HVS-enriched features. Consequently, this kind of knowledge cannot be embedded with the proposed method.

As future work, it would be interesting to investigate other forms of domain knowledge for building advanced BIQA models, such as image content and perceptual attributes (*e.g.* brightness, sharpness, and aesthetics). Since knowledge embedding is vital to the proposed framework, more effective embedding strategies are also worth studying.

References

[1] Z. Wang and A. C. Bovik, "Reduced- and no-reference image quality assessment," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 29–40, 2011.

[2] L. Chen, L. Wu, Z. Hu, and M. Wang, "Quality-aware unpaired image-to-image translation," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2664–2674, 2019.

[3] L. Li, Y. Zhou, J. Wu, F. Li, and G. Shi, "Quality index for view synthesis by measuring instance degradation and global appearance," *IEEE Transactions on Multimedia*, vol. 23, pp. 320–332, 2021.

[4] Q. Jiang, Z. Liu, K. Gu, F. Shao, X. Zhang, H. Liu, and W. Lin, "Single image super-resolution quality assessment: A real-world dataset, subjective studies, and an objective metric," *IEEE Transactions on Image Processing*, vol. 31, pp. 2279–2294, 2022.

[5] Q. Jiang, Y. Gu, C. Li, R. Cong, and F. Shao, "Underwater image enhancement quality evaluation: Benchmark dataset and objective metric," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 9, pp. 5959–5974, 2022.

[6] Q. Jiang, F. Shao, W. Lin, and G. Jiang, "Blique-tmi: Blind quality evaluator for tone-mapped images based on local and global feature analyses," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 29, no. 2, pp. 323–335, 2019.

[7] Q. Jiang, Z. Peng, F. Shao, K. Gu, Y. Zhang, W. Zhang, and W. Lin, "Stereoars: Quality evaluation for stereoscopic image retargeting with binocular inconsistency detection," *IEEE Transactions on Broadcasting*, vol. 68, no. 1, pp. 43–57, 2022.

[8] Z. Peng, Q. Jiang, F. Shao, W. Gao, and W. Lin, "Lggd+: Image retargeting quality assessment by measuring local and global geometric distortions," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 6, pp. 3422–3437, 2022.

[9] W. Chen, K. Gu, T. Zhao, G. Jiang, and P. L. Callet, "Semi-reference sonar image quality assessment based on task and visual perception," *IEEE Transactions on Multimedia*, vol. 23, pp. 1008–1020, 2021.

[10] Z. Wang, "Applications of objective image quality assessment methods [applications corner]," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 137–142, 2011.

[11] K. Gu, S. Wang, H. Yang, W. Lin, G. Zhai, X. Yang, and W. Zhang, "Saliency-guided quality assessment of screen content images," *IEEE Transactions on Multimedia*, vol. 18, no. 6, pp. 1098–1110, 2016.

[12] L. Zheng, L. Shen, J. Chen, P. An, and J. Luo, "No-reference quality assessment for screen content images based on hybrid region features fusion," *IEEE Transactions on Multimedia*, vol. 21, no. 8, pp. 2057–2070, 2019.

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

[14] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[15] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Using free energy principle for blind image quality assessment," *IEEE Transactions on Multimedia*, vol. 17, no. 1, pp. 50–63, 2014.

[16] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 1098–1105.

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

[18] P. Cui, S. Liu, and W. Zhu, "General knowledge embedded image representation learning," *IEEE Transactions on Multimedia*, vol. 20, no. 1, pp. 198–207, 2018.

[19] C. Chaudhary, P. Goyal, D. N. Prasad, and Y.-P. P. Chen, "Enhancing the quality of image tagging using a visio-textual knowledge base," *IEEE Transactions on Multimedia*, vol. 22, no. 4, pp. 897–911, 2020.

[20] C. Wei, H. Fan, S. Xie, C.-Y. Wu, A. Yuille, and C. Feichtenhofer, "Masked feature prediction for self-supervised visual pre-training," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 14 668–14 678.

[21] C. Tao, X. Zhu, G. Huang, Y. Qiao, X. Wang, and J. Dai, "Siamese image modeling for self-supervised vision representation learning," *arXiv preprint arXiv:2206.01204*, 2022.

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

[23] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 206–220, 2017.

[24] 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 Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2017.

[25] 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 Transactions on Image Processing*, vol. 27, no. 1, pp. 206–219, 2017.

[26] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2020.

[27] 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 *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3664–3673.

[28] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

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

[30] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 14 131–14 140.

[31] R. Ma, H. Luo, Q. Wu, K. N. Ngan, H. Li, F. Meng, and L. Xu, "Remember and reuse: Cross-task blind image quality assessment via relevance-aware incremental learning," in *Proceedings of the 29th ACM International Conference on Multimedia*, 2021, pp. 5248–5256.

[32] J. Liu, W. Zhou, J. Xu, X. Li, S. An, and Z. Chen, "Liqa: Lifelong blind image quality assessment," *arXiv preprint arXiv:2104.14115*, 2021.

[33] W. Zhang, D. Li, C. Ma, G. Zhai, X. Yang, and K. Ma, "Continual learning for blind image quality assessment," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE early access doi:10.1109/TPAMI.2022.3178874, 2022.

[34] D. R. Bull, "Chapter 2 - the human visual system," in *Communicating Pictures*. Oxford: Academic Press, 2014, pp. 17–61.

[35] J. Mannos and D. Sakrison, "The effects of a visual fidelity criterion of the encoding of images," *IEEE Transactions on Information Theory*, vol. 20, no. 4, pp. 525–536, 1974.

[36] J. Wu, F. Qi, and G. Shi, "Self-similarity based structural regularity for just noticeable difference estimation," *Journal of Visual Communication and Image Representation*, vol. 23, no. 6, p. 845–852, 2012.

[37] C.-H. Chou and Y.-C. Li, "A perceptually tuned subband image coder based on the measure of just-noticeable-distortion profile," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 5, no. 6, pp. 467–476, 1995.

[38] A. Srivastava, A. B. Lee, E. P. Simoncelli, and S.-C. Zhu, "On advances in statistical modeling of natural images," *Journal of Mathematical Imaging and Vision*, vol. 18, no. 1, pp. 17–33, 2003.

[39] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

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

[41] Q. Jiang, Z. Liu, S. Wang, F. Shao, and W. Lin, "Toward top-down just noticeable difference estimation of natural images," *IEEE Transactions on Image Processing*, vol. 31, pp. 3697–3712, 2022.

[42] I. J. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and harnessing adversarial examples," *arXiv preprint arXiv:1412.6572*, 2014.

[43] D. Taubman and M. Marcellin, "JPEG2000: standard for interactive imaging," *Proceedings of the IEEE*, vol. 90, no. 8, pp. 1336–1357, 2002.

[44] Y. Fang, H. Zhu, Y. Zeng, K. Ma, and Z. Wang, "Perceptual quality assessment of smartphone photography," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3674–3683.

[45] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2016.

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

[47] 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 Transactions on Image Processing*, vol. 20, no. 1, pp. 64–75, 2011.

[48] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, and L.-J. Li, "YFCC100M: The new data in multimedia research," *Communications of the ACM*, vol. 59, no. 2, p. 64–73, 2016.

[49] R. Caruana *et al.*, "Promoting poor features to supervisors: Some inputs work better as outputs," in *Advances in Neural Information Processing Systems*, 1996, pp. 389–395.

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

[51] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," *arXiv preprint arXiv: 2012.12877*, 2020.

[52] W. Zhang, K. Ma, G. Zhai, and X. Yang, "Task-specific normalization for continual learning of blind image quality models," *arXiv preprint arXiv:2107.13429*, 2021.

[53] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, and A. C. Bovik, "Image quality assessment using contrastive learning," *arXiv preprint arXiv:2110.13266*, 2021.

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

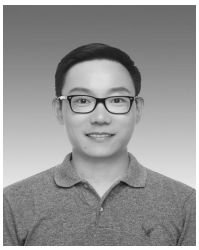
[55] D. Varga, D. Saupe, and T. Szirányi, "DeepRN: A content preserving deep architecture for blind image quality assessment," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2018, pp. 1–6.

[56] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal, Image and Video Processing*, vol. 12, no. 2, pp. 355–362, 2018.

[57] H. Lin, V. Hosu, and D. Saupe, "Kadid-10k: A large-scale artificially distorted iqa database," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 2019, pp. 1–3.

Tianshu Song received the B.S. degree in applied physics from China University of Mining and Technology, Xuzhou, China, in 2015, and the M.S. degree in electrical engineering from Shanghai University of Electric Power, Shanghai, China, in 2019. Currently, he is purchasing the Ph.D degree in the School of Information and Control Engineering, China University of Mining and Technology. His research interest is image quality assessment.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2022.3233244



Leida Li 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, Kaohsiung University of Science and Technology, Taiwan. From 2014 to 2015, he was a Visiting Research Fellow with the Rapid-Rich Object Search (ROSE) Laboratory, School of Electrical and Electronic Engineering, Nanyang Technological University, Singapore, where he was a Senior Research Fellow from 2016 to 2017. From 2009 to 2019, he worked in the School of Information and Control Engineering, China University of Mining and Technology, as Assistant Professor, Associate Professor and Professor, respectively. Currently, he is a Professor with the School of Artificial Intelligence, Xidian University.

His research interests include multimedia quality assessment, affective computing, information hiding, and image forensics. He has served as SPC for IJCAI 2019-2020, Session Chair for ICMR 2019 and PCM 2015, and TPC for AAAI 2019, ACM MM 2019-2020, ACM MM-Asia 2019, ACII 2019, PCM 2016. He is now an Associate Editor of the Journal of Visual Communication and Image Representation and the EURASIP Journal on Image and Video Processing.

Jinjian Wu received the B.S. and Ph.D. degrees from Xidian University, Xi'an, China, in 2008 and 2013, respectively. From 2011 to 2013, he was a Research Assistant with Nanyang Technological University, Singapore, where he was a Post-Doctoral Research Fellow from 2013 to 2014. From 2015 to 2019, he was an Associate Professor with Xidian University, where he had been a Professor since 2019. His research interests include visual perceptual modeling, biomimetic imaging, quality evaluation, and object detection. He received the Best Student Paper Award at the ISCAS 2013. He has served as associate editor for the journal of Circuits, Systems and Signal Processing (CSSP), the Special Section Chair for the IEEE Visual Communications and Image Processing (VCIP) 2017, and the Section Chair/Organizer/TPC member for the ICME 2014-2015, PCM 2015-2016, ICIP 2015, VCIP 2018, and AAAI 2019.

Yuzhe Yang 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.

Guangming Shi received the B.S. degree in automatic control, the M.S. degree in computer control, and the Ph.D. degree in electronic information technology from Xidian University, Xi'an, China, in 1985, 1988, and 2002, respectively. He had studied at the University of Illinois and University of Hong Kong. Since 2003, he has been a Professor with the School of Electronic Engineering, Xidian University. He was awarded the Cheung Kong Scholar Chair Professor by the Ministry of Education in 2012. He is currently the Academic Leader on circuits and systems, Xidian University. He has authored and coauthored more than 200 papers in journals and conferences.

His research interests include compressed sensing, brain cognition theory, multirate filter banks, image denoising, low-bitrate image and video coding, and implementation of algorithms for intelligent signal processing. He served as the Chair for the 90th MPEG and 50th JPEG of the international standards organization, and Technical Program Chair for FSKD06, VSPC 2009, IEEE Pulse Code Modulation 2009, SPIE Visual Communications and Image Processing 2010, and IEEE International Symposium on Circuits and Systems 2013.