This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2022.3179744, IEEE Transactions on Circuits and Systems for Video Technology

Blind Image Quality Assessment for Authentic Distortions by Intermediary Enhancement and Iterative Training

Tianshu Song, Leida Li, Pengfei Chen, Hantao Liu, Jiansheng Qian

***Abstract*—With the boom of deep neural networks, blind image quality assessment (BIQA) has achieved great processes. However, the current BIQA metrics are limited when evaluating low-quality images as compared to medium-quality and high-quality images, which restricts their applications in real world problems. In this paper, we first identify that two challenges caused by distribution shift and long-tailed distribution lead to the compromised performance on low-quality images. Then, we propose an intermediary enhancement-based bilateral network with iterative training strategy for solving these two challenges. Drawing on the experience of transitive transfer learning, the proposed metric adaptively introduces enhanced intermediary images to transfer more information to low-quality images for mitigating the distribution shift. Our metric also adopts an iterative training strategy to deal with the long-tailed distribution. This strategy decouples feature extraction and score regression for better representation learning and regressor training. It not only transfers the knowledge learned from the earlier stage to the latter stage, but also makes the model pay more attention to long-tailed low-quality images. We conduct extensive experiments on five authentically distorted image quality datasets. The results show that our metric significantly improves the evaluating performance on low-quality images and delivers state-of-the-art intra-dataset results. During generalization tests, our metric also achieves the best cross-dataset performance.**

***Index Terms*—image quality assessment, authentic distortion, low-quality, enhancement, generalization.**

I. Introduction

BLIND image quality assessment (BIQA) for authentic distortions has received intensive attention in recent years on account of its wide applications in many image processing fields, such as image capture, enhancement, retrieval, and transmission [1]–[11]. Since authentically distorted images do not have reference images and authentic distortions are much more complex than synthetic distortions, designing BIQA models for authentic distortions is challenging.

With the boom of deep neural networks, the current BIQA metrics for authentic distortions have achieved great advances.

This work was supported in part by the National Natural Science Foundation of China under Grants 62171340, 61991451 and 61771473, the Key Project of Shanxi Provincial Department of Education (Collaborative Innovation Center) under Grant 20JY024, and the Six Talent Peaks High-level Talents in Jiangsu Province under Grant XYDXX-063. (*Corresponding author: Leida Li*)

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

Leida Li and Pengfei Chen are with the School of Artificial Intelligence, Xidian University, Xi'an 710071, China (e-mails: ldli@xidian.edu.cn, cpf00790079@gmail.com).

Hantao Liu is with the School of Computer Science and Informatics, Cardiff University, Cardiff CF24 3AA, U.K. (e-mail: hantao.liu@cs.cardiff.ac.uk).

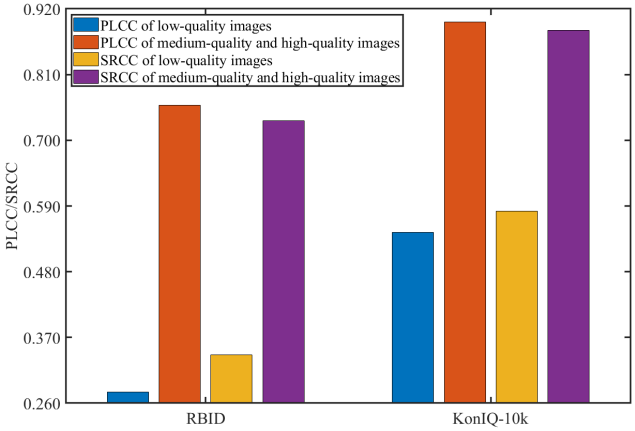

Fig. 1. Evaluation performance on images with different quality. The results are based on RBID [12] and KonIQ-10k [13] datasets. When image quality scores are labeled [12]–[14], the 25% quantile at the entire score interval is defined as the threshold of poor-quality. Following this definition, we define images with the lowest 25% scores as low-quality images, and other images are medium-quality and high-quality images.

However, the evaluation performance on low-quality images is significantly worse than that on medium-quality and high-quality images. To show the varying capacities on the evaluation of low-quality and medium-/high-quality images, we fine-tune a popular pre-trained Vision-Transformer model (data-efficient image transformer (DeiT) [15]) on two IQA datasets, and show the performance on images with different quality in Fig. 1 (some popular convolutional neural networks, *i.e.* ResNet18 [16] and VGG16 [17], achieving similar results). As observed from Fig. 1, the evaluation ability on low-quality images is rather limited.

Accurately assessing low-quality images is vital for IQA metrics because low-quality images widely exist in real life. For example, in image enhancement, the original images (even the enhanced images) are often with low-quality. To evaluate and improve the image enhancement algorithms, an ideal IQA model is expected to be able to precisely evaluate the qualities of original low-quality images and the enhanced images. Second, low-quality image assessment can help blind people. Many low-quality images are taken by blind people (*i.e.* VizWiz image quality issue dataset [18]), who try to overcome real visual challenges in their daily lives. Precisely evaluating those low-quality images and pointing out their flaws is helpful to obtain better images for blind photographers and further improve the VizWiz mobile application. Current

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2022.3179744, IEEE Transactions on Circuits and Systems for Video Technology

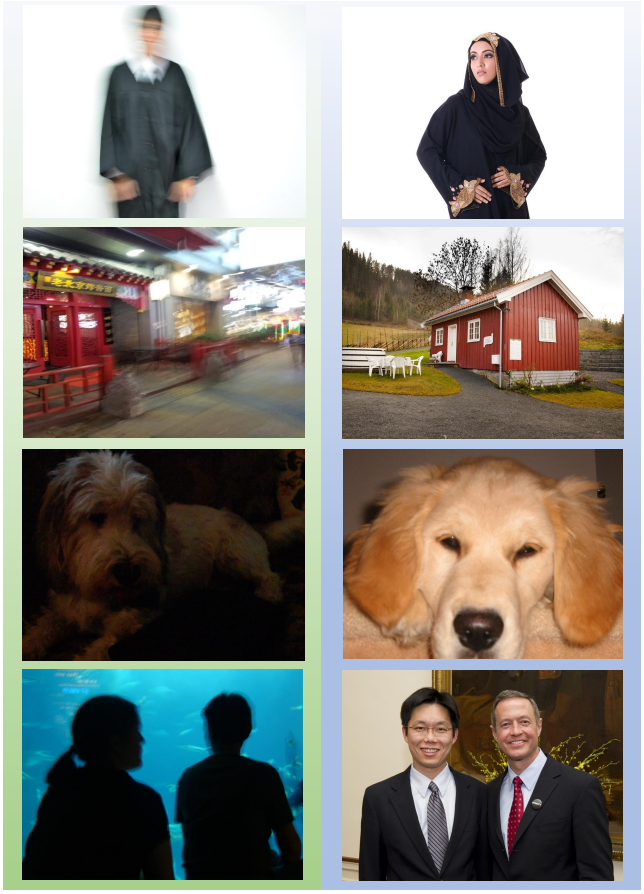

Fig. 2. Distribution shift between low-quality images and medium-/high-quality images. The left column shows low-quality images and the right column shows medium-/high-quality images.

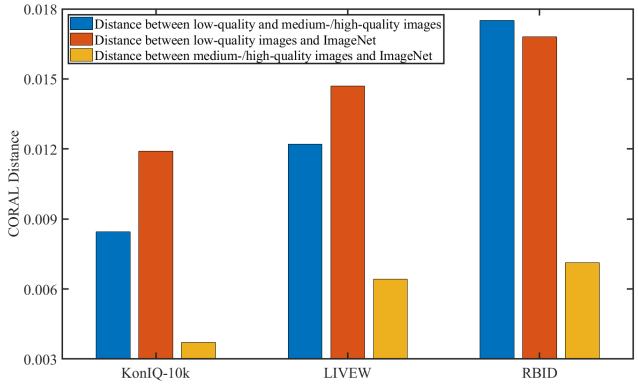

Fig. 3. Coral distance between low-quality images and medium-/high-quality images in IQA datasets and images in ImageNet.

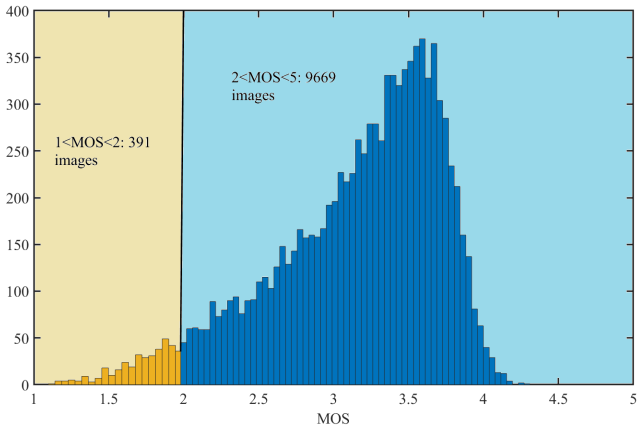

Fig. 4. Long-tailed distribution of mean opinion score (MOS) in KonIQ-10k. Low-quality images are marked with yellow, and other images are marked with blue. Low-quality images are images with the lowest 25% MOS.

IQA metrics fail to accurately evaluate low-quality images, limiting their real-world applications to a great extent.

The existing BIQA metrics rarely pay attention to such problem. Therefore, we start by identifying possible reasons. First, we deem that the distribution shift between low-quality images and medium-/high-quality images results in the limited evaluation ability. Most deep learning-based IQA metrics employ the pre-trained model on ImageNet [19] to mitigate the data shortage problem. Though this strategy improves model performance, it does not perform well on low-quality images. As shown in Fig. 2, low-quality images have noticeable differences from medium-/high-quality images, and they are quite different in distributions of brightness, contrast, sharpness, *etc*. Therefore, there exists a significant distribution shift between low-quality images and medium-/high-quality images. Since ImageNet mainly contains medium-quality and high-quality images, parameters trained on ImageNet cannot be transferred to IQA tasks for low-quality images satisfactorily. To intuitively show the distribution shift, following the method in [20], we adopt the CORAL distance [21] to measure 1) the distribution shift between low-quality images and medium-/high-quality images in IQA datasets, 2) the distribution shift between low-quality images in IQA datasets and images in ImageNet, 3) the distribution shift between medium-/high-quality images in IQA datasets and images in ImageNet. The results on three representative IQA datasets are shown in Fig.3, where larger distance means larger distribution shift. We can observe from Fig.3 that the distribution shift between medium-/high-quality images and images in ImageNet is the smallest. Low-quality images have much significant distribution shift from images in ImageNet and medium-/high-quality images. Second, the long-tailed distribution of IQA datasets also leads to the poor performance. Current authentically distorted IQA datasets typically have long-tailed score distributions, where low-quality images only account for a small portion of the whole dataset. We illustrate such phenomenon in Fig. 4, where the KonIQ-10k dataset [13] is shown considering its representativeness. It is known from Fig. 4 that the long-tailed distribution is apparent. Specifically, the total number of low-quality images (images with the lowest 25% quality scores [12]–[14]) account for only 3.9% of the whole dataset, which easily causes the trained model biased towards medium-/high-quality images.

Motivated by the above facts, this paper presents an intermediary enhancement-based bilateral network with iterative training strategy to deal with both challenges caused by distribution shift and imbalanced distribution. First, enlightened by transitive transfer learning [22], [23], the proposed

metric adaptively introduces an enhanced intermediary image for low-quality inputs to establish connection between low-quality images and high-quality images. The pre-trained model can extract more information from the enhanced intermediary image, which mitigates the distribution shift and strengthens the model's adaptability to low-quality images. Second, the proposed metric utilizes an iterative training strategy with two training stages to deal with the long-tailed distribution. This strategy decouples the feature extraction and score regression [24], [25]. By this means, the proposed method transfers the knowledge learned at the first stage for targeted training of low-quality images at the second stage. It also makes the model pay more attention to low-quality images which are prone to be ignored due to their long-tailed distribution. Therefore, our method alleviates the challenge caused by imbalanced distribution. Finally, with the bilateral network, the proposed method integrates a conventional feature extractor and a low-quality image enhanced feature extractor, which significantly improves the evaluation ability on low-quality images while guaranteeing the evaluation ability on common images.

The contributions of this paper are summarized as follows:

- We propose a new BIQA model with intermediary enhancement and iterative training for enhancing the evaluation ability on low-quality images.
- We propose an approach to build a connection between low-quality images and high-quality images by adaptively introducing enhanced intermediary images to strengthen the pre-trained model's adaptability to low-quality images, which mitigates the distribution shift.
- We introduce an iterative training strategy to tackle the long-tailed distribution problem of low-quality images. This makes the model pay more attention to low-quality images, which further improves the evaluation ability.

The rest of this paper is arranged as follows. In Section II, we briefly review the existing BIQA metrics, image enhancement algorithms, and long-tailed learning methods. The proposed method is detailed in Section III, and experiments are presented in Section IV. Finally, Section V concludes this paper.

II. Related Work

A. Blind Image Quality Assessment

Traditional BIQA metrics typically utilized a score regressor with handcrafted features (such as BRISQUE [26] and NFERM [27]), or codebooks (such as CORNIA [28] and HOSA [29]) for evaluating image quality. However, handcrafted features or codebooks are limited in describing authentic distortions.

With the boom of deep learning, deep neural networks have been widely adopted to extract features for authentically distorted images. Earlier attempts tried to extract features from small patches with specially designed shallow neural networks [30]–[33]. Though these metrics achieved decent performance on synthetic distortions, their performance on authentic distortions are unsatisfactory due to the small sample property of IQA and the limited feature representation ability of shallow networks. To avoid these two issues, many recent BIQA models employ ImageNet pre-trained network for extracting quality-aware features [34]–[39]. For example, Zhu *et al.* [38] fine-tuned models pre-trained on ImageNet with synthetic distortions to extract the meta-knowledge for the subsequent fine-tuning on authentic distortions. Zhang *et al.* [34] adopted a two-stream network, called DBCNN, for quality evaluation. The backbone of one stream was pre-trained with many synthetically distorted images, and the other stream was pre-trained on ImageNet. Following this framework, Zhang *et al.* [35] proposed a new approach, called UNIQUE, to train the two-stream network with a mixed dataset obtained from six IQA datasets with both synthetic distortions and authentic distortions. Some recent IQA metrics adopted lifelong strategies to improve the generalization ability of IQA models. For example, Liu *et al.* [40] proposed a lifelong IQA metric through a split-and-merge distillation strategy. Zhang *et al.* [41] continually trained an IQA model on a stream of IQA datasets by adopting the continual learning strategy. Ma *et al.* [42] proposed a remember-and-reuse network to perform the cross-task IQA based on the incremental learning strategy. Though these metrics have achieved great process for evaluating authentic distortions, we observe that they are still limited in evaluating low-quality images due to the following reasons. First, as discussed above, models pre-trained on ImageNet (mainly consisting of medium-/high-quality images) cannot easily transfer to low-quality images due to the distribution shift. Second, the long-tailed distribution property of IQA datasets causes those models biased towards medium-/high-quality images. Therefore, we propose a new IQA model for mitigating challenges of distribution shift and long-tailed distribution.

B. Image Enhancement

The proposed metric introduces enhanced images for mitigating the distribution shift. Since low-light and blur are most common in low-quality images, low-light image enhancement (LLIE) and image deblurring are briefly reviewed. Many popular LLIE algorithms [43] are based on the Retinex theory [44], which assumes that the observed color image can be decomposed into reflectance and illumination. RetinexNet [45] is the representative of Retinex theory-based metrics, which successfully combined deep neural networks with Retinex theory. It first decomposed an low-light image into two parts of reflectance and illumination, and then enhanced the decomposed images. Finally, it fused the decomposed images and obtained the enhanced image with visually pleasing quality.

Many popular deep learning-based image deblurring algorithms, such as SRN-DeblurNet [46] and DeblurGAN [47], have been proposed for synthetically blurred images. For example, to deblur images from coarse-to-fine, SRN-DeblurNet specially designed a scale-recurrent network for processing multi-scale inputs. To achieve more image details during deblurring, DeblurGAN adopted generative adversarial networks (GANs) to deblur images, because GANs are adept at generating vivid details. To deblur real-world blur distortions, Rim *et al.* [48] first created a real blur-based dataset, containing image

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2022.3179744, IEEE Transactions on Circuits and Systems for Video Technology

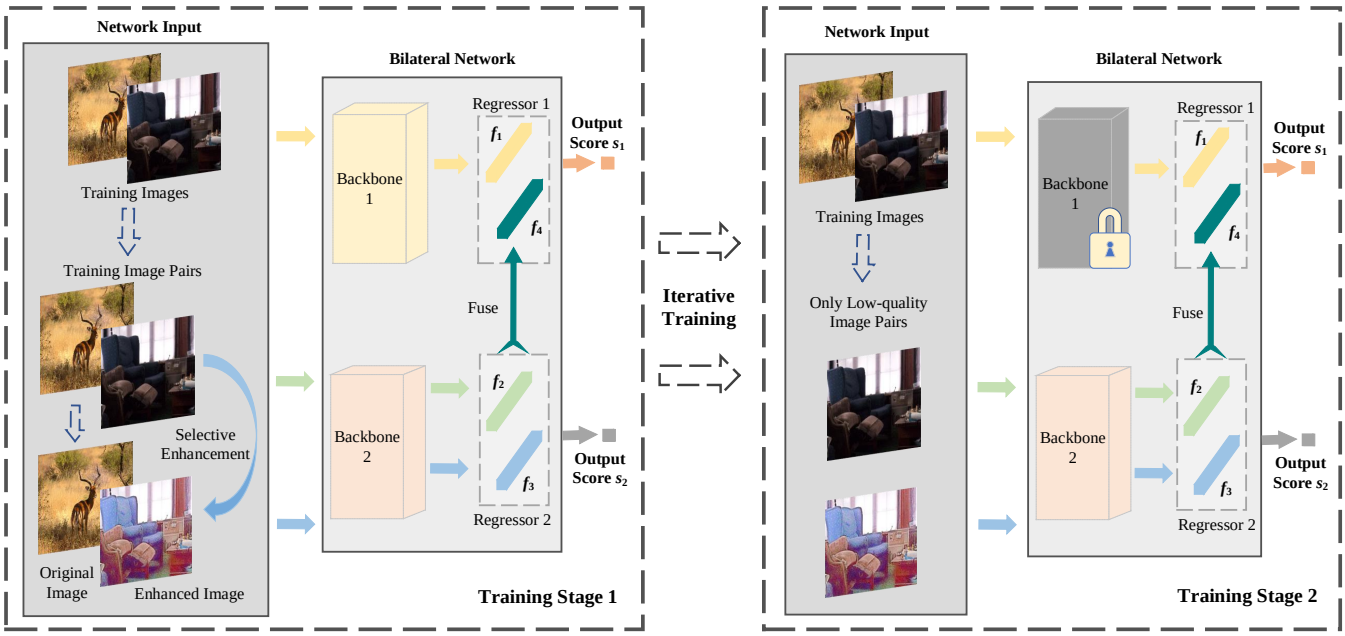

Fig. 5. The proposed blind image quality assessment by Intermediary Enhancement and Iterative Training (IEIT). The framework consists of two training stages based on a bilateral network. The input of 'Backbone 1' is a batch of images, and the input of 'Backbone 2' is a batch of image pairs. The predicted score s_1 of 'Regressor 1' at the second stage is the final image quality score. During the second training stage, 'Backbone 1' is frozen and 'Backbone 2' is trained with low-quality image pairs only.

pairs of real-wolrd blurred images and the corresponding ground-truth sharp images. Then, they trained the above two models on this dataset. Models trained on this dataset achieved much better performance on authentic blur distortions, which is adopted to enhance blur images in this paper.

C. Deep Long-tailed Learning

One challenge leading to the poor evaluation ability on low-quality images is the long-tailed distribution property of IQA datasets. Therefore, deep long-tailed learning [24] is another research field related to this paper, which includes some popular strategies of resampling [49], reweighting [50] and decoupled training [51]. Resampling strategy aims to balance the number of training sample of different classes during model training. Zhang *et al.* [49] adopted a specially designed loss to determine the sampling rate for different classes, aiming to sample more training data for the under-represented tail classes. The reweighting strategy typically re-balances classes by adjusting loss. Lin *et al.* [50] proposed focal-loss with higher weights for the hard-to-learn tail classes and lower weights for the easy-to-learn head classes. Decoupled training is designed for better representation learning and classifier training by decoupling the training procedure into feature extraction and classifier training. Chu *et al.* [51] first trained feature extractor, and then made augmentation on tail-classes for re-training the classifier.

Most of the current deep long-tailed learning approaches focus on classification tasks or detection/segmentation tasks, but rare works pay attention to regression tasks. As stated in [52], the long-tailed distribution for the regression task has different properties from the classification task, and which strategy is effective for IQA task remains largely unknown. In this paper, we make attempts to mitigate the problem of long-tailed distribution on low-quality images.

III. Proposed Method

In this paper, we propose a new BIQA model via Intermediary Enhancement and Iterative Training (IEIT) with the objective to enhance the evaluation ability of low-quality images, which has been mainly caused by the distribution shift and long-tailed distribution. Specifically, to improve the evaluation ability on low-quality images and guarantee the evaluation ability on other images, the proposed method adopts a bilateral network to integrate a conventional feature extractor and a low-quality image enhanced feature extractor. With the objective of mitigating the distribution shift and strengthening the feature extractor's adaptability to low-quality images, we introduce the enhanced images for the low-quality image enhanced feature extractor to extract more information about low-quality images. For alleviating the long-tailed distribution challenge and making the model pay more attention to low-quality images, we utilize an iterative training strategy to transfer the knowledge learned at the first stage for targeted training of low-quality images at the second stage. The whole structure of the proposed method is illustrated in Fig. 5, which consists of a bilateral network with enhanced input images and two training stages. The top backbone is a conventional feature extractor, whose input is a batch of images, aiming to learn conventional quality assessment features. The bottom backbone is a low-quality image strengthened feature extractor, whose input is a batch of image pairs composed of original images and the corresponding selectively enhanced images. To improve the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2022.3179744, IEEE Transactions on Circuits and Systems for Video Technology

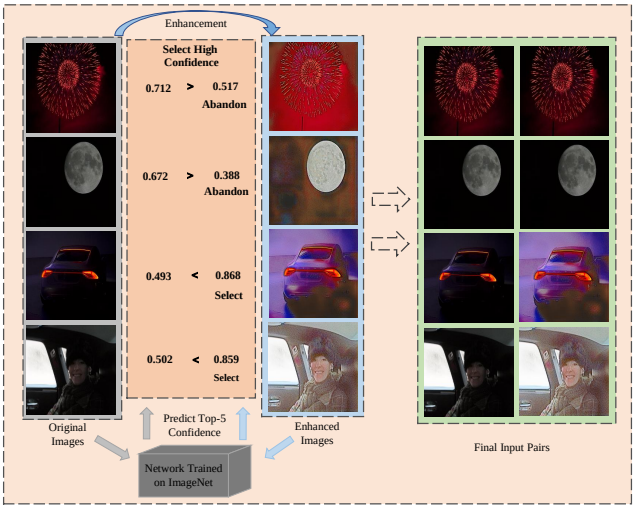

Fig. 6. The process of selective enhancement. Enhanced images with higher prediction confidence are chosen, and enhanced images with lower confidence are replaced with their original images.

evaluation ability on low-quality images while guaranteeing the ability on common images, the bilateral network integrates both feature extractors, and make predictions by combining both features through the regressor in the top lateral.

A. Selective Enhancement

To mitigate the distribution shift, the proposed method introduces intermediary enhanced images. The pre-trained model can extract more useful information from enhanced images, which improves its transferability to low-quality images. However, one may doubt that some enhanced images may even be harmful to the adaptability, because current enhancement algorithms are still not flawless. To address this issue, we propose a selective strategy. For our approach, it is vital to make the pre-trained model understand enhanced images. If the pre-trained model has difficulty in understanding the enhanced images, the extracted features will consequently have low reliability and practicability, and vice versa. Since the pre-trained model is a classification model, better-understood images tend to achieve higher classification accuracy. Due to the lack of semantic labels in IQA datasets, we utilize recognition confidence as an alternative, which is related to the understanding of an image [53]. Based on this, we compare the top-5 prediction probability of original images with that of enhanced images. To ensure that enhanced images are better understood, only enhanced images with higher prediction confidences are chosen, and enhanced images with lower confidence are replaced with their original images directly. By this means, better-understood images are introduced, which provide more information and mitigates the distribution shift. The whole process of selective enhancement is illustrated in Fig. 6.

In our method, we make selective enhancement on images with low-light and blur distortions. These two kinds of distortions widely exist in low-quality images, such as the Smartphone Photography Attribute and Quality (SPAQ) dataset [39], which includes labels of image quality, sharpness and brightness. For intuitive understanding, we show the histogram distributions of brightness and sharpness of images with different quality in Fig. 7. From Fig. 7 we can observe that compared with medium-quality and high-quality images, low-quality images tend to have lower brightness and sharpness. Therefore, we adopt LLIE and image deblurring algorithms with the proposed selective strategy to enhance images for tackling the distribution shift challenge.

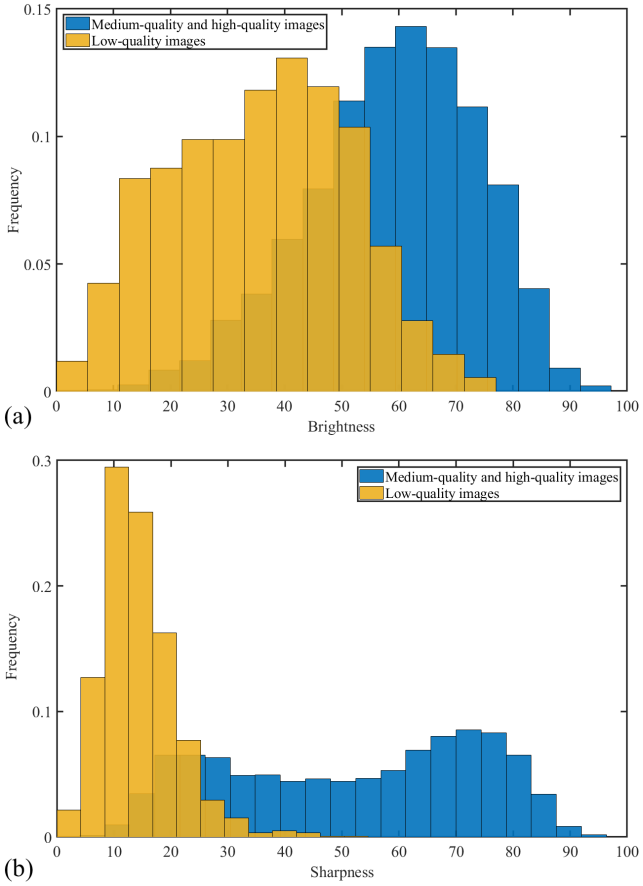

Fig. 7. The histogram of brightness and sharpness of low-quality images and medium-quality and good-quality images in SPAQ dataset. (a) shows the histogram distributions of brightness. (b) shows the histogram distributions of sharpness.

B. Iterative Training

We adopt an iterative training strategy with two training stages to overcome the challenge caused by long-tailed distribution. During the first training stage, the top backbone, denoted as V_1, extracts conventional quality assessment features $\boldsymbol{f}_1$ from a batch of images $\boldsymbol{I}$. The bottom backbone, denoted as $\boldsymbol{V}_2$, extracts features $\boldsymbol{f}_2$ and $\boldsymbol{f}_3$ from a batch of image pairs $(\boldsymbol{I}, \boldsymbol{E})$ composed of original images and the corresponding selective enhanced images, where $\boldsymbol{f}_2, \boldsymbol{f}_3$ corresponds to $\boldsymbol{I}, \boldsymbol{E}$ respectively. Then, the 'Regressor 1', denoted as R_1, predicts the quality score s_1 from feature $\boldsymbol{f}_1$ and $\boldsymbol{f}_4$, where $\boldsymbol{f}_4$ is fused by $\boldsymbol{f}_2$ and $\boldsymbol{f}_3$. Next, the 'Regressor 2', denoted as R_2, predicts the quality score s_2 from $\boldsymbol{f}_2$ and $\boldsymbol{f}_3$. Finally, these two scores are sent to the loss function for back-propagation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2022.3179744, IEEE Transactions on Circuits and Systems for Video Technology

Algorithm 1. Training details of the first stage.

Inputs: A batch of images $\boldsymbol{I}$ for V_1; a batch of image pairs $(\boldsymbol{I}, \boldsymbol{E})$ for V_2; 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 $\boldsymbol{f}_1$ from $\boldsymbol{I}$:
$\boldsymbol{f}_1 = V_1(\boldsymbol{I})$;
// Obtain $\boldsymbol{f}_2, \boldsymbol{f}_3$ from $(\boldsymbol{I}, \boldsymbol{E})$:
$\boldsymbol{f}_2, \boldsymbol{f}_3 = V_2(\boldsymbol{I}, \boldsymbol{E})$;
// Obtain $\boldsymbol{f}_4$ by fusing $\boldsymbol{f}_2, \boldsymbol{f}_3$:
$\boldsymbol{f}_4 = F(\boldsymbol{f}_2, \boldsymbol{f}_3)$;
// Obtain s_1 from $\boldsymbol{f}_1, \boldsymbol{f}_4$:
$s_1 = R_1(\boldsymbol{f}_1, \boldsymbol{f}_4)$;
//Obtain s_2 from $\boldsymbol{f}_2, \boldsymbol{f}_3$:
$s_2 = R_2(\boldsymbol{f}_2, \boldsymbol{f}_3)$;
//Obtain the loss of Regressor 1, denoted as l_1:
$l_1 = \frac{1}{n}\sum_{i=1}^{n} abs(y_i - s_{1,i}), i = 1, 2, ..., N$;
//Obtain the loss of Regressor 2, denoted as l_2:
$l_2 = \frac{1}{n}\sum_{i=1}^{n} abs(y_i - s_{2,i}), i = 1, 2, ..., N$;
//Obtain the total loss l_t from l_1, l_2:
$l_t = l_1 + \lambda l_2$;

Return: l_t.

to train V_1, V_2, R_1, R_2. The loss function we utilized is mean absolute error (MAE). The whole process is summarized in Algorithm 1.

One objective of the first stage is to train the backbone V_1 for extracting conventional quality assessment features, that are adept at describing medium-/high-quality images. The other objective is make the backbone V_2 acquire image quality related knowledge for the target training at the second stage. After the first training stage, both objectives are achieved.

During the second training stage, the inputs of both backbones are still images and image pairs. The differences between these two stages are that parameters of V_1 are frozen during the second training stage. Parameters of V_2, R_2 are trained by low-quality image pairs only. At this stage, medium-/high-quality image pairs having no contribution to V_2, R_2, they only affect parameters of R_1.

In specific, to train V_2, R_2, we first extract features $\boldsymbol{f}_2, \boldsymbol{f}_3$ from low-quality image pairs $(\boldsymbol{I}_l, \boldsymbol{E}_l)$, and then predicts s_2 from them:

$$\begin{cases} \boldsymbol{f}_2, \boldsymbol{f}_3 = V_2(\boldsymbol{I}_l, \boldsymbol{E}_l), \\ s_2 = R_2(\boldsymbol{f}_2, \boldsymbol{f}_3). \end{cases} \tag{1}$$

Next, the predicted score s_2 is sent into the MAE loss function for back-propagation:

$$l_2 = \frac{1}{n}\sum_{j=1}^{n} abs(y_j - s_{2,j}), j = 1, 2, ..., N, y_j < t; \tag{2}$$

where y is MOS, n is the batch size of training images, and t is the threshold value for low-quality images, which means only low-quality image pairs contribute to l_2.

After V_2 is optimized with low-quality images, we finally fine-tune R_1 with all images. In specific, we first freeze the backbone V_1, and then extract conventional quality assessment features $\boldsymbol{f}_1$:

$$\boldsymbol{f}_1 = d(V_1(\boldsymbol{I})) \ , \tag{3}$$

where the function '$d(\cdot)$' means detaching the gradients at the inference stage. This function it will stop the gradients during back-propagation.

To train R_1, we also need $\boldsymbol{f}_4$, which is obtained by fusing output features $\boldsymbol{f}_2, \boldsymbol{f}_3$:

$$\begin{cases} \boldsymbol{f}_2, \boldsymbol{f}_3 = d(V_2(\boldsymbol{I}, \boldsymbol{E})) \ , \\ \boldsymbol{f}_4 = F(\boldsymbol{f}_2, \boldsymbol{f}_3). \end{cases} \tag{4}$$

Then, $\boldsymbol{f}_4$ is fused with $\boldsymbol{f}_1$ for predicting score s_1:

$$s_1 = R_1(\boldsymbol{f}_1, \boldsymbol{f}_4). \tag{5}$$

Finally, the predicted score s_1 is sent into the MAE loss function for training R_1:

$$l_1 = \frac{1}{n}\sum_{i=1}^{n} abs(y_i - s_{1,i}), i = 1, 2, ..., N. \tag{6}$$

At the second training stage, V_1 pre-trained on the first stage is just utilized for extracting conventional quality assessment features $\boldsymbol{f}_1$. The quality related knowledge in V_2 learned at the first stage is transferred to the second stage, which is helpful for learning better representations for low-quality images. After fine-tuning with low-quality images, V_2 has paid more attention to low-quality images, which mitigates the challenge that the model biases towards medium-/high-quality images due to the long-tailed distribution. During this stage, R_1 predicts from the combination of conventional quality assessment features and features optimized for low-quality images, which improves the evaluation performance for low-quality images and ensures performance on common images.

At the test stage, the inputs are the same as the first training stage. Image $\boldsymbol{I}$ and its enhanced image $\boldsymbol{E}$, V_1, V_2, R_1 are necessary, and R_2 is abandoned. In specific, we first obtain conventional quality assessment features $\boldsymbol{f}_1$ from backbone V_1:

$$\boldsymbol{f}_1 = V_1(\boldsymbol{I}). \tag{7}$$

Then we obtain low-quality image strengthened feature $\boldsymbol{f}_4$ from low-quality image strengthened backbone V_2:

$$\begin{cases} \boldsymbol{f}_2, \boldsymbol{f}_3 = V_2(\boldsymbol{I}, \boldsymbol{E}); \\ \boldsymbol{f}_4 = F(\boldsymbol{f}_2, \boldsymbol{f}_3). \end{cases} \tag{8}$$

Finally, we predict the final image quality score s_1 by combining conventional quality assessment feature and low-quality image strengthened feature:

$$s_1 = R_1(\boldsymbol{f}_1, \boldsymbol{f}_4). \tag{9}$$

C. Network Training

Different image regions have varying effects on the quality of the entire image. Therefore, mining the relationship between different image regions and global content is vital to the IQA task [55]. However, one basic principle of convolutional neural network (CNN) is adopting small receptive fields, which leads to the difficulty in mining the relationship between faraway regions. In contrast, Vision-Transformer (ViT) [56] utilizes pure multiheaded self-attention (MSA) modules [57] in processing vision tasks. The MSA module is good at processing

TABLE I
DETAILED INFORMATION OF FIVE AUTHENTIC IQA DATASETS.

Dataset	KonIQ-10k [13]	SPAQ [39]	LIVEW [14]	CID2013 [54]	RBID [12]
Number of Images	10,073	11,125	1162	480	585
Quality Score Type	MOS	MOS	MOS	MOS	MOS
Quality Score 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
Annotators	1459	N/A	8100	188	180
Total Rating Number	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	Selected from YFCC100m	Captured with 66 mobile phones in the wild	Captured with 15 digital devices in the wild	Captured with 79 devices in eight scenes	Captured with 1 digital camera in the wild

long-range contextual information and global information. Therefore, ViT can mine the relationship between different image regions more comprehensively, which benefits the IQA task. Consequently, we choose the Vision-Transformer as the backbone of our model.

The regressors we adopted are three fully connected (FC) layers, whose neuron numbers decrease from the dimension of fused features to one. Each FC layer is followed by a rectified linear unit (RELU) as the activation function. The output of the last FC layer is the final quality score. During training, this score is then fed into the MAE loss function for back-propagation. The optimizer we adopted is the stochastic gradient decrease (SGD) with warming-up strategy.

IV. EXPERIMENTS

A. Evaluation Protocol and Implementation Details

In this paper, we conduct experiments on five authentically distorted image quality datasets, including KonIQ-10k dataset [13], Smartphone Photography Attribute and Quality dataset (SPAQ) [39], LIVE in the Wild Image Quality Challenge (LIVEW) dataset [14], RBID dataset [12] and CID2013 dataset [54]. The detailed information is summarized in Table. I.

In all experiments, we adopt the most prevalent Pearson Linear Correlation Coefficient (PLCC) and Spearman Linear Correlation Coefficient (SRCC) as evaluation criteria. For the predicted score sequence $\{s_1, s_2, ..., s_n\}$ and the target label sequence $\{y_1, y_2, ..., y_n\}$, PLCC and SRCC can be calculated by:

$$\mathrm{PLCC} = \frac{n\sum s_i y_i - \sum s_i \sum y_i}{\sqrt{n\sum s_i^2 - (\sum s_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}, \quad (10)$$

$$\mathrm{SRCC} = 1 - \frac{6\sum d_i^2}{n(n^2-1)}, \quad (11)$$

where d_i means the difference of ranks of two sequences.

In this paper, we utilize the backbone of data-efficient image transformer (DeiT) [15] as our feature extractor. DeiT has the same structure as ViT, and it solves the weakness of ViT, which needs too many images for training (300 million images in JFT300M [64]). The DeiT can be trained with much less resources and achieves better performance than ViT. In specific, we adopt DeiT-Small pre-trained on ImageNet with totally 22M parameters, which is less than widely used models of ResNet50 (26M parameters) and VGG19 (144M parameters). Therefore, it can well converge during training. On our experimental setting (NVIDIA Titan-XP GPU), it can infer 35 images per second with batch size of 1 and 2200 images with batch size of 64. It has an input size of 224×224, with 12 blocks and 6-headed self attention. Each image is split into 14×14 patches, and each patch is mapped to a 384-dimensional vector. During training, we first resize the original images into 244×244, and then randomly crop images of size 224×224 as input.

For image enhancement, we choose DeblurGAN-v2 [65] trained on real-world blur dataset [48] to deblur images and RetinexNet [45] to enhance the low-light images. Before regressing the quality score, we need three fusion operations, which fuses $\boldsymbol{f}_2, \boldsymbol{f}_3$ for obtaining s_2 and $\boldsymbol{f}_4$, and fuses $\boldsymbol{f}_1, \boldsymbol{f}_4$ to achieve s_1. In order to reduce the complexity and training difficulty, we only tried several simple fusing strategies. We tried the operation of concatenation, minus, multiplication for obtaining $\boldsymbol{f}_4$, and concatenation, add, multiplication for s_1, s_4. Finally, concatenation achieves the best performance. Therefore, we utilize the concatenation for feature fusion.

During the first training stage, the learning rate is 0.03 with the warming-up strategy (warming-up with 0.001 and 0.005 for 10 epochs respectively). And the learning rate at the second training stage is 0.01 (warming-up with 0.001 and 0.003 for 10 epochs respectively). When the performance does not grow up for 20 epochs, the learning rate is multiplied by 0.3 until the learning rate is smaller than 5×10^{-5}. The parameter λ in loss function of Algorithm 1 is 1. In our experiments, we randomly split the dataset into training (80%) and test (20%) subsets for 10 times, and then the average performance across all repetitions on the test subset is reported.

B. Performance Evaluation

The primary objective of the proposed model is to improve the evaluation ability on low-quality images. Therefore, we first conduct the intra-dataset experiments on low-quality images. Low-quality images in one dataset are defined as images with the lowest 25% MOS [12]–[14]. Since rare researches focus on low-quality images, performance on low-quality images are rarely reported. Therefore, to make comparisons, we retrain some popular BIQA methods, including handcrafted feature-based metrics such as NFERM [27], BRISQUE [26] and HOSA [29], and code-available deep learning-based metrics of WaDIQaM-NR [33], UNQIUE [35], DBCNN [34]

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2022.3179744, IEEE Transactions on Circuits and Systems for Video Technology

TABLE II
PLCC/SRCC RESULTS OF LOW-QUALITY IMAGES. UNIQUE ONLY ACHIEVES RESULTS ON THREE DATASETS DUE TO ITS SPECIALLY DESIGNED MIXED TRAINING DATASET.

Dataset	KonIQ-10k [13]		SPAQ [39]		LIVEW [14]		CID2013 [54]		RBID [12]		Weighted Average	
Criteria	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
NFERM [27]	0.086	0.075	0.331	0.315	0.319	0.327	0.133	0.152	0.199	0.205	0.218	0.206
BRISQUE [26]	0.090	0.103	0.286	0.298	0.079	0.078	0.057	0.058	0.127	0.141	0.183	0.194
CORNIA [28]	0.280	0.284	0.307	0.329	0.228	0.212	0.338	0.319	0.196	0.216	0.289	0.300
HOSA [29]	0.226	0.249	0.321	0.325	0.166	0.148	0.191	0.188	0.179	0.182	0.266	0.277
BMPRI [58]	0.021	0.063	0.048	0.029	-0.232	-0.216	-0.326	-0.393	-0.100	-0.141	0.011	0.019
SCORER [59]	0.122	0.134	0.222	0.233	0.223	0.200	0.221	0.224	0.209	0.212	0.179	0.188
SNP-NIQE [60]	-0.106	-0.118	-0.256	-0.191	-0.041	-0.081	-0.337	-0.556	-0.120	-0.200	-0.179	-0.162
QUEADI [61]	0.082	0.091	0.267	0.271	0.113	0.078	0.064	0.033	0.062	0.036	0.171	0.173
UNIQUE [35]	0.347	0.338	/	/	0.590	0.605	/	/	0.295	0.313	0.368	0.363
WaDIQaM-NR [33]	0.021	0.048	0.439	0.411	0.224	0.192	0.492	0.474	0.194	0.133	0.244	0.238
DBCNN [34]	0.436	0.300	0.481	0.445	0.306	0.376	0.571	0.559	0.159	0.167	0.447	0.375
MetaIQA [38]	0.448	0.476	0.465	0.425	0.579	0.552	0.671	0.683	0.133	0.098	0.459	0.450
IEIT	**0.597**	**0.611**	**0.498**	**0.470**	**0.663**	**0.681**	**0.691**	**0.694**	**0.392**	**0.378**	**0.550**	**0.543**

TABLE III
PLCC/SRCC RESULTS OF INTRA-DATASET TESTS.

Dataset	KonIQ-10k [13]		SPAQ [39]		LIVEW [14]		CID2013 [54]		RBID [12]		Weighted Average	
Criteria	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
NFERM [27]	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 [26]	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 [28]	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 [29]	0.791	0.761	0.873	0.866	0.703	0.667	0.835	0.833	0.716	0.684	0.825	0.806
BMPRI [58]	0.152	0.175	-0.436	-0.399	-0.218	-0.141	-0.035	0.155	-0.159	-0.113	-0.121	-0.157
SCORER [59]	0.762	0.732	0.831	0.827	0.619	0.608	0.838	0.835	0.670	0.654	0.787	0.771
SNP-NIQE [60]	0.069	0.116	-0.500	-0.507	-0.296	-0.289	-0.170	0.011	-0.281	-0.288	-0.233	-0.212
QUEADI [61]	0.737	0.711	0.837	0.835	0.594	0.565	0.752	0.758	0.601	0.578	0.774	0.760
NSSADNN [62]	/	/	/	/	0.813	0.745	0.825	0.748	/	/	0.817	0.746
MEON [63]	/	/	/	/	0.693	0.688	0.703	0.701	/	/	0.696	0.692
BIECON [32]	/	/	/	/	0.613	0.595	0.620	0.606	/	/	0.615	0.598
Zhang *et al.* 2021 [37]	/	0.847	/	/	/	0.835	/	/	/	0.827	/	0.845
UNIQUE [35]	0.901	**0.896**	/	/	**0.890**	**0.854**	/	/	**0.873**	**0.858**	0.899	0.890
CONTRIQUE [36]	0.906	0.894	0.919	0.914	0.857	0.845	/	/	/	/	0.910	**0.901**
WaDIQaM-NR [33]	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 [34]	0.869	0.856	0.915	0.911	0.869	0.851	0.871	0.863	0.859	0.845	0.891	0.882
MetaIQA [38]	0.887	0.850	0.871	0.870	0.835	0.802	0.784	0.766	0.777	0.746	0.872	0.853
IEIT	**0.916**	0.892	**0.921**	**0.917**	0.865	0.833	**0.891**	**0.874**	0.839	0.809	**0.913**	0.899

and MetaIQA [38]. Different from other metrics, UNIQUE is originally designed to be trained with a special mixed-dataset [35], consisting of KonIQ-10k, LIVEW, RBID, and other three synthetic IQA datasets. Therefore, only results on these three authentic datasets are achieved. The performance on low-quality images of all methods are listed in Table II, and we can observe from Table II that the evaluation ability of current popular BIQA metrics are very limited, and the proposed metric consistently outperforms all other metrics on all five databases by a large margin.

An ideal IQA metric should not only work well for low-quality images but also has decent overall performance. Therefore, we further summarize the results on the entire dataset in Table III, which consists of eight metrics in Table II (such as DBCNN, UNIQUE), and other five deep neural network-based metrics with reported results on authentic IQA datasets. Results in Table III show that the proposed metric also achieves SOTA performance during the whole dataset tests, especially on the two largest datasets of KonIQ-10k and SPAQ. In Table III, UNIQUE and DBCNN achieve better results on some small datasets because they were specially pre-trained with many synthetically distorted images, and then were fine-tuned on target IQA datasets (UNIQUE was fine-tuned with six datasets). Benefiting from much more training images, they achieve better performance than our method on small datasets. To make the results more intuitive, we show the scatter plots in Fig. 8. As observed from Fig. 8, the points of the proposed metric are more densely clustered around the red line.

C. Cross-Dataset Tests

In addition to the intra-dataset test, we also test the generalization ability of the proposed model based on cross-dataset validation, which is vital in real-world applications. Since the KonIQ-10k dataset has the largest image capacity, we train BIQA models on KonIQ-10k and then directly test them on other datasets.

Different datasets adopt different strategies for labeling image quality. Directly utilizing models trained on one dataset to assess other datasets may introduce intrinsic deviation

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2022.3179744, IEEE Transactions on Circuits and Systems for Video Technology

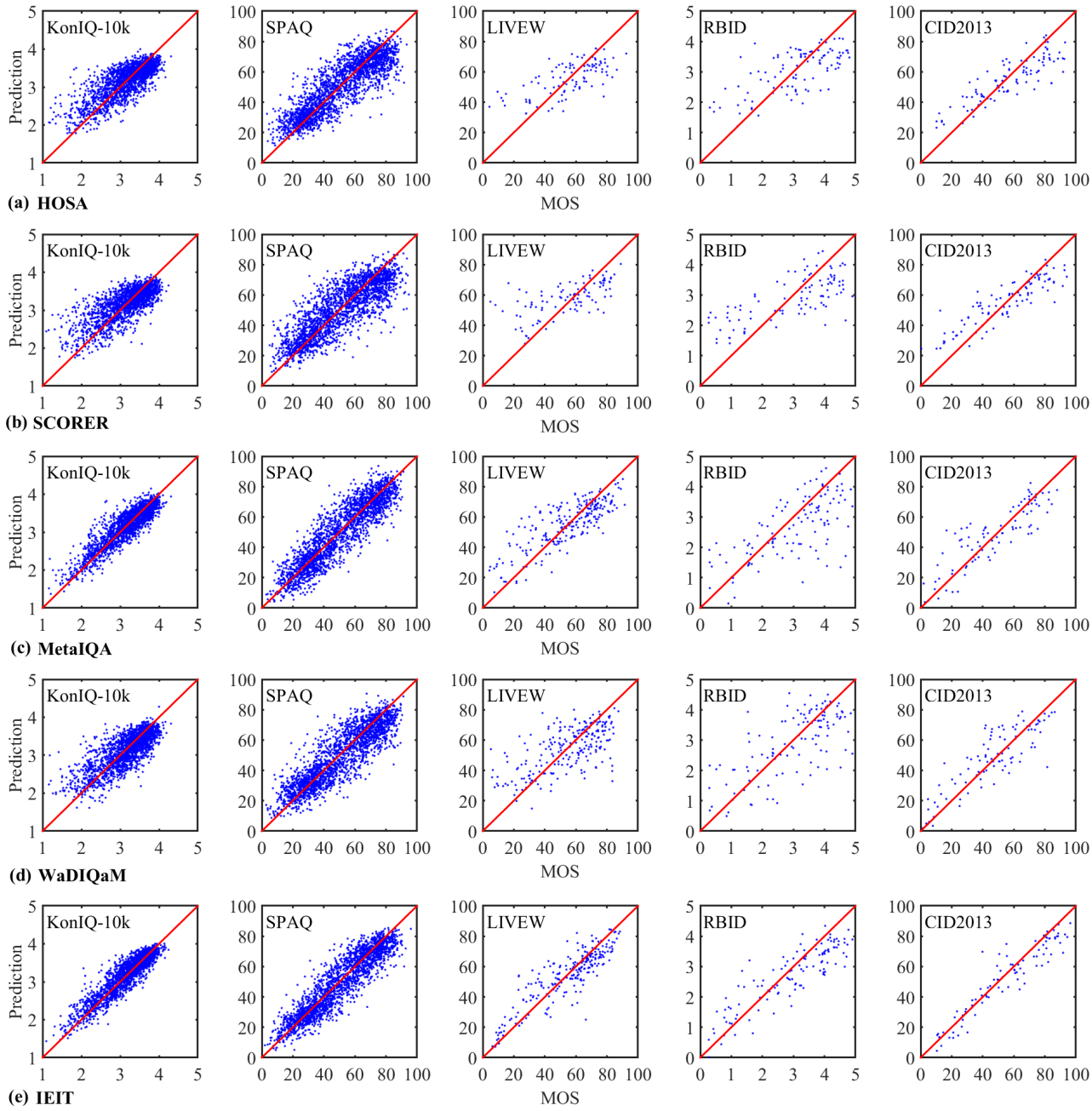

Fig. 8. The scatter plots of intra-dataset tests. From top to bottom, each line shows results of HOSA, SCORER, MetaIQA, WaDIQaM, and IEIT, respectively. From left to right, each column shows results from the dataset of KonIQ-10k, SPAQ, LIVEW, RBID, and CID2013, respectively.

[68]. Therefore, before calculating PLCC/SRCC during cross-dataset tests, we first adopt a nonlinear-mapping function to map the predicted scores. As suggested by the Video Quality Experts Group (VQEG2000) [69], we can map the relative quality score to the perceptual quality score by a four parameter nonlinear function:

$$Q_p = f(Q_r) = \frac{\beta_3 - \beta_4}{1 + e^{-\frac{Q_r - \beta_1}{|\beta_2|}}} + \beta_4 \tag{12}$$

where Q_r, Q_p are the relative quality score and the perceptual quality score respectively. We can observe from Eq. (12) that this mapping does not affect the final results of SRCC, and β_3, β_4 do not affect the results of PLCC.

Following the method in [68], MOS ranges in all datasets are linearly rescaled to the value range of 1-5 first, which means $\beta_3 = 5, \beta_4 = 1$. Then, we construct a model with one middle output score Q_r followed by five mapping head to learn the parameters of $\beta_j^i, j = 1, 2, i = 1, 2, 3, 4, 5$ ($\beta_3^i = 5, \beta_4^i = 1$). Next, we train the model on all datasets, and each head corresponds to one dataset. Finally, we obtain the five groups of parameters β_j^i, and each i corresponds to one dataset.

After obtaining those parameters, we can map the output scores predicted by the model trained on KonIQ-10k for

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2022.3179744, IEEE Transactions on Circuits and Systems for Video Technology

TABLE IV
PLCC/SRCC RESULTS OF CROSS-DATASET TEST. THE MODEL IS TRAINED ON KONIQ-10K AND DIRECTLY TESTED ON OTHER DATASETS.

Dataset	SPAQ [39]		LIVEW [14]		CID2013 [54]		RBID [12]		Weighted Average	
Criteria	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
NFERM [27]	0.688	0.711	0.548	0.540	0.715	0.680	0.520	0.530	0.669	0.687
BRISQUE [26]	0.650	0.682	0.575	0.554	0.555	0.533	0.581	0.597	0.637	0.662
CORNIA [28]	0.711	0.766	0.672	0.639	0.605	0.538	0.686	0.688	0.703	0.743
HOSA [29]	0.731	0.771	0.675	0.652	0.690	0.664	0.692	0.679	0.723	0.753
SCORER [59]	0.663	0.588	0.554	0.536	0.654	0.635	0.515	0.518	0.647	0.582
QUEADI [61]	0.143	0.116	0.050	0.082	-0.043	-0.061	0.083	0.101	0.126	0.106
DeepRN (ResNet101) [66]	/	/	0.750	0.726	/	/	/	/	0.750	0.726
DeepBIQ (InceptionV2) [67]	/	/	0.821	0.804	/	/	/	/	0.821	0.804
ConCept512 [13]	/	/	0.848	0.825	/	/	/	/	0.848	0.825
UNIQUE [35]	/	/	/	0.786	/	/	/	0.783	/	0.785
WaDIQaM-NR [33]	0.743	0.779	0.653	0.647	0.702	0.676	0.629	0.659	0.729	0.759
DBCNN [34]	0.851	0.850	0.764	0.729	0.781	0.736	0.777	0.784	0.838	0.833
MetaIQA [38]	0.834	0.851	0.806	0.783	0.764	0.710	0.780	0.781	0.827	0.837
IEIT	**0.868**	**0.870**	**0.858**	**0.829**	**0.816**	**0.771**	**0.831**	**0.822**	**0.864**	**0.861**

TABLE V
PLCC/SRCC RESULTS OF ABLATION STUDIES. 'IE' MEANS INTERMEDIARY ENHANCEMENT. 'IT' MEANS ITERATIVE TRAINING.

Dataset	KonIQ-10k [13]		SPAQ [39]		LIVEW [14]		CID2013 [54]		RBID [12]	
Criteria	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
Baseline	0.546	0.581	0.452	0.436	0.594	0.655	0.611	0.656	0.295	0.338
w/ IE (w/o IT)	0.553	0.583	0.476	0.458	0.624	0.658	0.645	0.668	0.353	0.350
w/ IT (w/o IE)	0.577	0.593	0.481	0.458	0.631	0.648	0.684	0.693	0.360	0.371
w/ IE+IT	**0.597**	**0.611**	**0.498**	**0.470**	**0.663**	**0.681**	**0.691**	**0.694**	**0.392**	**0.378**

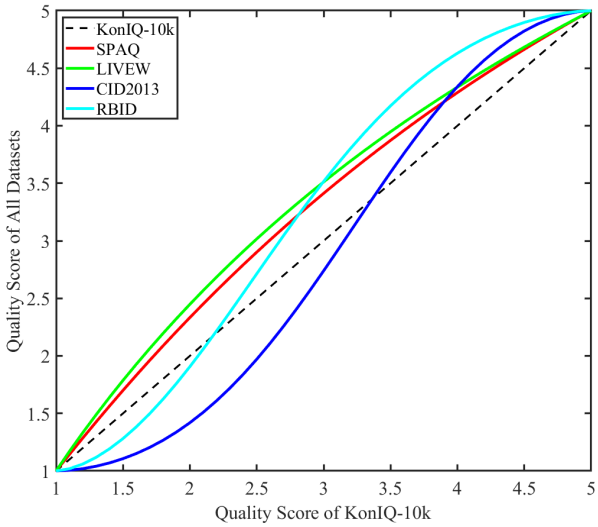

Fig. 9. Nonlinear mapping functions from KonIQ-10k to other datasets.

predicting other four datasets:

$$\begin{cases} Q_r = f^{-1}(Q_p^i) = \beta_1^i - ln(\frac{\beta_3^i - Q_p^i}{Q_p^i - \beta_4^i})\left|\beta_2^i\right|, i = 1; \\ Q_p^i = f(Q_r) = \frac{\beta_3^i - \beta_4^i}{1+e^{-\frac{Q_r-\beta_1^i}{|\beta_2^i|}}} + \beta_4^i, i = 2,3,4,5; \end{cases} \tag{13}$$

where $i = 1$ represents parameters of the KonIQ-10k dataset, and $i = 2, 3, 4, 5$ represents parameters of other four datasets. The mapping functions from KonIQ-10k to other datasets are shown in Fig. 9

After adopting the nonlinear mapping, the cross-dataset evaluation results are shown in Table IV. Table IV clearly shows that our framework achieves the best generalization performance on all test datasets. Results on LIVEW and RBID are even comparable to the intra-dataset evaluation scenario. The high generalization performance of the proposed metric owes to the better evaluation ability on low-quality images and the valuable information provided by the enhanced images.

D. Ablation Studies

The proposed method introduces enhanced intermediary images for mitigating the distribution shift. It also takes an iterative training strategy to address the long-tailed distribution challenge. To prove the effectiveness of the enhanced intermediary image and iterative training strategy, we make ablation studies.

First, we start from a baseline model which is pre-trained on ImageNet. We then train it on IQA databases without intermediary enhancement and iterative training, the results are listed in the first row of Table V. Then, we introduce the intermediary enhancement without the iterative training. Next, we introduce the iterative training without the intermediary enhancement. Finally, we list the results of the proposed method with both intermediary enhancement and iterative training in the last row of Table V.

Comparing the second row with the first row in Table V, we can observe that results trained with intermediary enhancement are significantly better than results of the baseline model. Because the enhanced image builds a connection between low-quality images and high-quality images, which provides more useful information and mitigates the distribution shift. The performance on low-quality images is consequently improved.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TCSVT.2022.3179744, IEEE Transactions on Circuits and Systems for Video Technology

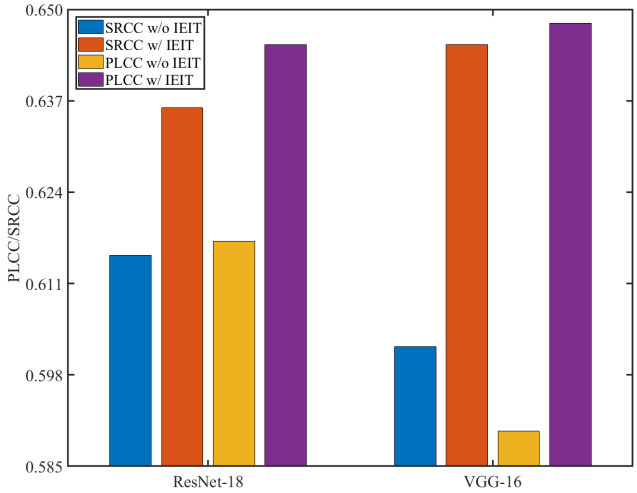

Fig. 10. Ablation results on ResNet18 and VGG16. The reported results are based on the LIVEW dataset.

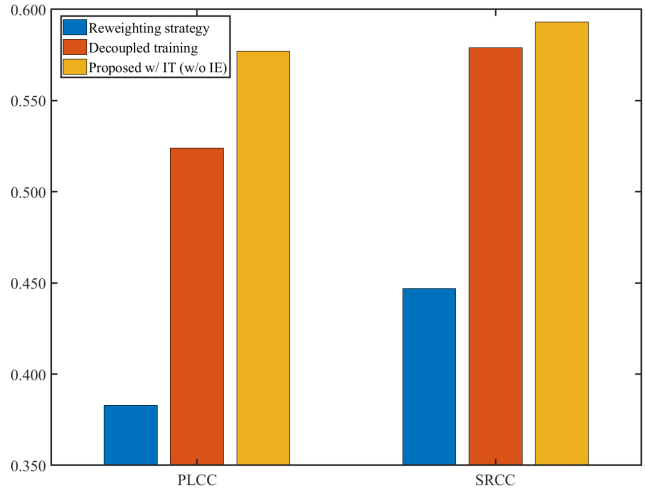

Fig. 11. Comparison of different long-tailed deep learning strategies. The reported results are based on the KonIQ-10k dataset.

Adopting iterative training strategy achieves better results than the baseline model, revealing that iterative training strategy is efficient for evaluating low-quality images as well. Since the knowledge learned from the first stage is transferred to the second stage and the model pays more attention to low-quality images during the second training stage, the challenge of model biasing towards medium-/high-quality images caused by long-tailed distribution is mitigated, which leads to the better performance.

Finally, the proposed metric introduces both enhanced images and iterative training strategy, which overcomes both challenges posed by distribution shift and long-tailed distribution. Consequently, the proposed model with both components delivers the best performance.

The proposed metric adopts the image enhancement and iterative training strategies, which do not depend on a specific network structure and can be generalized to popular convolutional neural networks (CNNs). For example, we take experiments with ResNet18 and VGG16, and show the results in Fig.10. It can be observed from Fig.10 that the proposed IEIT can significantly improve the evaluation ability of CNNs.

E. Evaluation of Other Long-tailed Learning Strategies

As aforementioned, deep long-tailed learning for regression is rarely investigated. We also make some other attempts by directly adopting the reweighting strategy and simply taking the decoupled training strategy. For the reweighting strategy, we increase the weight of low-quality images in the loss function to make the model pay more attention to low-quality images. For the decoupled training strategy, we first train the backbone and regressor together, and then fine-tune the regressor only (freeze the backbone) by adopting the above reweighting strategy. Finally, for fair comparisons with the above deep long-tailed learning methods, we also show the results of proposed method without the intermediary enhancement (same as the third row in Table V) in Fig. 11.

Fig. 11 shows that directly adopting the reweighting strategy achieves the worst performance, and the decoupled training strategy achieves better performance. The performance of the proposed method is the best. Here are some possible reasons. Simply utilizing reweighting strategy during training may be not conducive to the feature representation for medium-/high-quality images, and the knowledge learned from medium-/high-quality images is also vital for evaluating low-quality images. Therefore, decoupled training strategy achieves better performance by conventionally learning the feature representation and fine-tuning the regressor only. The proposed method not only keeps conventional quality assessment features but also combines it with low-quality image strengthened features. Therefore, it achieves the best results, which also means the proposed method apparently mitigating the problem of long-tailed distribution.

V. Conclusions

In this paper, we have proposed a novel BIQA model, with the objective to enhance the evaluation ability on low-quality images. The proposed model not only adaptively introduces an intermediary enhanced image to mitigate the distribution shift challenge, but also adopts an iterative training strategy for solving the long-tailed distribution challenge. Extensive experimental results show that the proposed model significantly improves evaluation ability on low-quality images and achieves SOTA intra-dataset results. The proposed metric also obtains the best generalization performance during cross-dataset tests.

References

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

[2] X. Zhang, S. Wang, K. Gu, W. Lin, S. Ma, and W. Gao, "Just-noticeable difference-based perceptual optimization for jpeg compression," *IEEE Signal Processing Letters*, vol. 24, no. 1, pp. 96–100, 2017.

[3] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "Generalizable no-reference image quality assessment via deep meta-learning," *IEEE Transactions on Circuits and Systems for Video Technology*, IEEE early access doi: 10.1109/TCSVT.2021.3073410, 2021.

[4] Y. Fang, R. Du, Y. Zuo, W. Wen, and L. Li, "Perceptual quality assessment for screen content images by spatial continuity," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 4050–4063, 2020.

[5] S. Wang, K. Gu, X. Zhang, W. Lin, S. Ma, and W. Gao, "Reduced-reference quality assessment of screen content images," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 1, pp. 1–14, 2018.

[6] Q. Wu, L. Wang, K. N. Ngan, H. Li, F. Meng, and L. Xu, "Subjective and objective de-raining quality assessment towards authentic rain image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 11, pp. 3883–3897, 2020.

[7] 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.

[8] H. G. Kim, H.-T. Lim, and Y. M. Ro, "Deep virtual reality image quality assessment with human perception guider for omnidirectional image," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 917–928, 2020.

[9] Y. Zhou, Y. Sun, L. Li, K. Gu, and Y. Fang, "Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network," *IEEE Transactions on Circuits and Systems for Video Technology*, IEEE early access, doi: 10.1109/TCSVT.2021.3073410, 2021.

[10] Y. Niu, H. Zhang, W. Guo, and R. Ji, "Image quality assessment for color correction based on color contrast similarity and color value difference," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 4, pp. 849–862, 2018.

[11] Q. Wu, H. Li, K. N. Ngan, and K. Ma, "Blind image quality assessment using local consistency aware retriever and uncertainty aware evaluator," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 9, pp. 2078–2089, 2018.

[12] 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.

[13] 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.

[14] 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.

[15] 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.

[16] 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) (CVPR)*, 2016, pp. 770–778.

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

[18] T.-Y. Chiu, Y. Zhao, and D. Gurari, "Assessing image quality issues for real-world problems," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 3643–3653.

[19] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Feifei, "ImageNet large scale visual recognition challenge," *International Journal of Computer Vision*, vol. 115, no. 3, pp. 211–252, 2015.

[20] B. Li, W. Zhang, M. Tian, G. Zhai, and X. Wang, "Blindly assess quality of in-the-wild videos via quality-aware pre-training and motion perception," *IEEE Transactions on Circuits and Systems for Video Technology*, 2022.

[21] B. Sun and K. Saenko, "Deep coral: Correlation alignment for deep domain adaptation," in *European conference on computer vision*. Springer, 2016, pp. 443–450.

[22] B. Tan, Y. Song, E. Zhong, and Q. Yang, "Transitive transfer learning," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2015, pp. 1155–1164.

[23] B. Tan, Y. Zhang, S. Pan, and Q. Yang, "Distant domain transfer learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, 2017, pp. 2604–2610.

[24] Y. Zhang, B. Kang, B. Hooi, S. Yan, and J. Feng, "Deep long-tailed learning: A survey," *arXiv preprint arXiv:2110.04596*, 2021.

[25] B. Kang, S. Xie, M. Rohrbach, Z. Yan, A. Gordo, J. Feng, and Y. Kalantidis, "Decoupling representation and classifier for long-tailed recognition," in *International Conference on Learning Representations (ICLR)*, 2020, pp. 1–16.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 1733–1740.

[31] L. Kang, P. Ye, Y. Li, and D. Doermann, "Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks," in *IEEE International Conference on Image Processing (ICIP)*, 2015, pp. 2791–2795.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] 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.

[41] W. Zhang, D. Li, C. Ma, G. Zhai, X. Yang, and K. Ma, "Continual learning for blind image quality assessment," *arXiv preprint arXiv:2102.09717*, 2021.

[42] 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.

[43] C. Li, C. Guo, L.-H. Han, J. Jiang, M.-M. Cheng, J. Gu, and C. C. Loy, "Low-light image and video enhancement using deep learning: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, IEEE early access doi: 10.1109/TPAMI.2021.3126387, 2021.

[44] E. H. Land, "The retinex theory of color vision," *Scientific American*, vol. 237, no. 6, pp. 108–129, 1977.

[45] C. Wei, W. Wang, W. Yang, and J. Liu, "Deep retinex decomposition for low-light enhancement," in *British Machine Vision Conference*, 2018, pp. 1–11.

[46] X. Tao, H. Gao, X. Shen, J. Wang, and J. Jia, "Scale-recurrent network for deep image deblurring," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8174–8182.

[47] O. Kupyn, V. Budzan, M. Mykhailych, D. Mishkin, and J. Matas, "DeblurGAN: Blind motion deblurring using conditional adversarial networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 8183–8192.

[48] J. Rim, H. Lee, J. Won, and S. Cho, "Real-world blur dataset for learning and benchmarking deblurring algorithms," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 184–201.

[49] Y. Zang, C. Huang, and C. C. Loy, "FASA: Feature augmentation and sampling adaptation for long-tailed instance segmentation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 3457–3466.

[50] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 2980–2988.

[51] P. Chu, X. Bian, S. Liu, and H. Ling, "Feature space augmentation for long-tailed data," in *European Conference on Computer Vision (ECCV)*, 2020, pp. 694–710.

[52] Y. Yang, K. Zha, Y.-C. Chen, H. Wang, and D. Katabi, "Delving into deep imbalanced regression," in *International Conference on Machine Learning (ICML)*, 2021, pp. 11 842–11 851.

[53] T. Song, L. Li, H. Zhu, and J. Qian, "Ie-iqa: Intelligibility enriched generalizable no-reference image quality assessment," *Frontiers in Neuroscience*, vol. 15, pp. 1–12, 2021.

[54] 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.

[55] 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 Transactions on Circuits and Systems for Video Technology*, IEEE early access doi: 10.1109/TCSVT.2021.3112197, 2021.

[56] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby, "An image is worth 16x16 words: Transformers for image recognition at scale," *arXiv preprint arXiv: 2010.11929*, 2020.

[57] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in Neural Information Processing Systems (NIPS)*, 2017, pp. 1–11.

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

[59] M. Oszust, "Local feature descriptor and derivative filters for blind image quality assessment," *IEEE Signal Processing Letters*, vol. 26, no. 2, pp. 322–326, 2019.

[60] Y. Liu, K. Gu, Y. Zhang, X. Li, G. Zhai, D. Zhao, and W. Gao, "Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 929–943, 2020.

[61] M. Rajchel and M. Oszust, "No-reference image quality assessment of authentically distorted images with global and local statistics," *Signal, Image and Video Processing*, vol. 15, no. 1, pp. 83–91, 2021.

[62] 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.

[63] 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.

[64] C. Sun, A. Shrivastava, S. Singh, and A. Gupta, "Revisiting unreasonable effectiveness of data in deep learning era," in *IEEE International Conference on Computer Vision (ICCV)*, 2017, pp. 843–852.

[65] O. Kupyn, T. Martyniuk, J. Wu, and Z. Wang, "DeblurGAN-v2: Deblurring (orders-of-magnitude) faster and better," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 8878–8887.

[66] 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.

[67] 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.

[68] D. Li, T. Jiang, and M. Jiang, "Unified quality assessment of in-the-wild videos with mixed datasets training," *International Journal of Computer Vision*, vol. 129, no. 4, pp. 1238–1257, 2021.

[69] V. Q. E. Group *et al.*, "Final report from the video quality experts group on the validation of objective models of video quality assessment," in *VQEG meeting*, 2000, pp. 1–129.

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.

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.

Pengfei Chen received the B.S. degree from Xidian University, Xi'an, China, in 2014 and the Ph.D. degree from China University of Mining and Technology, Xuzhou, China, in 2022. He is currently a lecturer with the School of Artificial Intelligence, Xidian University. His research interests include image/video quality assessment, video quality of experience, and domain adaptation/generalization.

Hantao Liu received the Ph.D. degree from the Delft University of Technology, Delft, The Netherlands, in 2011. He is currently an Associate Professor with the School of Computer Science and Informatics, Cardiff University, Cardiff, U.K. He is an Associate Editor of the IEEE Transactions on Human Machine Systems and the IEEE Transactions on Multimedia.

Jiansheng Qian received the B.S. degree from Xidian University, Xi'an, China, in 1985, the M.S. degree from Xi'an Institute of Optics and Precision Mechanics, Chinese Academy of Sciences, Xi'an, China, in 1988, and the Ph.D. degree in 2003. He is currently a Professor with the China University of Mining and Technology, Xuzhou, China. His research interests include the areas of signal processing for communication, broadband network technology and applications, and coal mine communication and monitoring.