REQA: Coarse-to-fine Assessment of Image Quality to Alleviate the Range Effect
Bingheng Li, and Fushuo Huo

Abstract—Blind image quality assessment (BIQA) of User Generated Content (UGC) suffers from the range effect, which indicates that on the overall quality range, mean opinion score (MOS) and predicted MOS (pMOS) are well correlated while focusing on a particular range, the correlation is lower. To tackle this problem, a novel method is proposed from coarse-grained metric to fine-grained prediction. Concretely, we utilize global context features and local detailed features for the multi-scale distortion perception. Then, to further boost the ability of fine-grained assessment, we introduce the feedback mechanism, which is in accord with Human Vision System (HVS), to perceive detailed distortions gradually. Also, two coarse-to-fine loss functions are proposed to facilitate the feedback perception progress: a rank-and-gradient loss for coarse-grained metric keeps the assessment rank and gradient consistency between pMOS and MOS; a multi-level tolerance loss following the curriculum learning strategy is proposed to make a fine-grained prediction. Both coarse-grained and fine-grained experiments demonstrate that the proposed method outperforms the state-of-the-art ones, which validates that our method effectively alleviates the range effect. The codes are available at https://github.com/huofushuo/REQA.

Index Terms—Blind image quality assessment, range effect, coarse-to-fine assessment, feedback hierarchy

I. INTRODUCTION

I mage quality assessment (IQA) explores how to imitate human beings to automatically assess image quality. Accurately describing the quality change has extensive applications in image restoration [1], image compression [2], point cloud processing [3], etc. IQA approaches can be generally divided into three categories: full-reference IQA approach (FR-IQA), reduced-reference approach (RR-IQA), and blind IQA approach (BIQA). FR-IQA and RR-IQA measure the similarity between the distorted image and reference image [4–9]. However, in most authentic scenarios, it is hard to achieve ideal reference information. BIQA does not require any reference image as a prerequisite in predicting perceptual quality, so it has attracted great attention in recent years.

Early BIQA methods [12–24] mainly focus on synthetically distorted databases [25–29], which consist of multiple distortions generated from limited scenarios. Since images captured in the real world suffer from ever-changing contents and more complicated distortions, accurately predicting their quality remains a challenge. To deal with this problem, in recent years, some BIQA methods [10, 30–40] have been proposed, exploring novel learning strategies and leveraging complicated feature representation. However, although achieving relatively high correlations spanning a wide range from extremely bad to good on challenge datasets [11, 14, 22], these approaches may be confronted with a drawback that they cannot perform well on a narrow quality range, which is referred as the range effect [43–45]. Fig. 1 shows a visual example of the range effect of HyperIQA [10] on the CLIVE [11] dataset.

How well BIQA methods perform on a narrow quality range is important for a realistic prediction of User Generated Content (UGC). According to [44, 46], the distributions of image quality in the authentically distorted databases are narrow and peak as compared to those synthetic ones, as most pictures captured in the real world are improved by the imaging device. According to the distribution property, the streaming media servers can screen the images with extreme qualities to obtain images with moderate qualities. These images will be integrated as the benchmark of subsequent image enhancement and image-effect synthesis. Thus, we claim that the BIQA model should possess more a fine-grained perception ability to assess image quality, alleviating the range effect.

Some methods [44, 45] proposed a new evaluation criterion aiming at FR-IQA, which eliminates the range effect to some extent. However, there are few attempts to solve this problem in BIQA yet. Recently, Zhang et al. [47] revisited
the IQA and conducted a survey on the fine-grained IQA. They pointed out that existing IQA methods do not address the potential fine-grained IQA. In this paper, we first attempt to develop a more fine-grained IQA assessment method to alleviate the Range Effect, termed as REQA. Our method follows the coarse-to-fine principle. For the coarse-grained assessment, existing metric-learning-based methods map the samples into the multi-dimensional feature vectors, as for clustering or nearest neighbor classification utilizing a distance function that captures pair-wised similarity. This is effective on synthetic images. These methods map the synthetically distorted images into the feature vectors which represent distortion classification or distortion level. However, in authentic databases, images possess nonidentical scenes and immeasurable distortions that increase the difficulty to obtain the coherence of feature distributions. Even if it can be achieved between a pair of images, it may not be quality-aware. The proposed method maps a set of images to supervised scores, based on these scores, a rank-and-gradient metric can be conducted as for reducing the prediction deviation from a wide quality rank and further alleviate the order confusion of the predicted quality sequence. For the fine-grained assessment, which is unsolved for existing methods, we adopt a multi-stage prediction strategy. Previous BIQA approaches adopt a one-time strategy to predict MOS with a feedforward structure. However, the way neglects the feedback mechanism of the Human Visual System (HVS) as it is important for perceptual learning. Neurologically speaking, feedforward hierarchy underlies implicit processing for initial vision at a glance, and feedback connections add details to explicit vision with scrutiny. The same applies to BIQA where fine-grained cognition of image quality is achieved through feedback processing where high-level and low-level features are recurrently integrated by HVS. Moreover, the feedback-based learning approach has been proven more effective than the commonly employed feedforward paradigm in prediction tasks. In this paper, MOS prediction is constantly refined under multi-level tolerance constraints through a feedback structure. Besides, the coarse-grained metric is fused into the structure as the prior knowledge as it is easier compared to fine-grained prediction from the perspective of curriculum learning.

In summary, our contributions are four-fold:

- To our best knowledge, we take the first attempt to develop a fine-grained blind image quality assessment method to alleviate the range effect.
- We propose the effective coarse-to-fine strategy, which not only utilizes global context features and local detailed features for the multi-scale distortion perception but also introduces the feedback mechanism to perceive detailed distortions gradually.
- We also devise two coarse-to-fine loss functions to facilitate the feedback perception progress.
- Comprehensive experiments based on traditional coarse-grained evaluation and fine-grained evaluation show the effectiveness of our method.

II. Related Work

In this section, as we try to handle the range effect in BIQA, we give the detailed review of BIQA methods for synthetically distorted images and authentically distorted images, respectively.

A. BIQA for Synthetically Distorted Images

Previous research of BIQA approaches for the synthetic task mainly follows two kinds of ideas: traditional methods and learning-based methods.

Commonly traditional BIQA approaches first extract hand-crafted features based on the empirical analysis and then adopt a regression function to map the features into the quality score. The most well-known category is based on the characteristics of natural statistical scenes (NSS), such as DIIVINE [54], BLINDS-II [55] and BRISQUE [56]. This type of method assumes that natural images have certain statistical properties affected by distortion, which would make the image look unnatural. Therefore, features can be extracted from frequency, spatial, and wavelet domains based on the statistical properties of an image to predict its quality score. In addition, there are also some other methods based on the human visual system (HVS), such as NRSL [57] and RISE [58]. These methods utilize HVS to construct the quality-aware features, assuming that HVS is adapted to the structure information. However, these hand-crafted features are time-consuming and meantime lack of generalization ability due to the diversity of image contents and distortions.

Unlike the traditional BIQA methods, the learning-based BIQA approaches automatically generate quality-aware features. In the early stage, CBIQ [59] and CORNIA [12] introduced the code-book feature-based learning into BIQA. These methods first utilize raw image patches extracted from a set of unlabeled images to learn a dictionary in an unsupervised manner, and then encode the test images on the dictionary to obtain the feature representations for quality estimation. In the subsequent development, CNN pushes the significant development of BIQA thanks to its powerful learning ability. WaDIQaM [14] proposes a significantly deeper framework that comprises ten convolutional layers and five pooling layers for feature extraction, and two fully connected layers for regression. As these networks grow deeper and wider, they need larger annotated databases for training. Due to extremely labor-intensive and costly subjective experiment, the current IQA databases are too small to meet this requirement.

To deal with the small sample problem, it needs to explore more effective learning strategies and leverage more complicated features. Some methods find a way out in transfer learning. MEON [15] and DBCNN [30] pretrain a classification model on a large-scale synthetic database to acquire the initialized network parameters which are, to some extent, distortion-aware. RankIQA [24] and dipIQ [16] propose a pairwise learning-to-rank (L2R) algorithm, which can learn to rank images in terms of image distortion. Then, they transfer the prior knowledge learned from ranked images to a traditional CNN. After fine-tuning it on IQA databases, it can improve the accuracy of IQA. RRLRIQA [17] makes
further exploration and models the BIQA as a Markov decision process to optimize the whole image-quality ordering directly. During training phase, not all distortions or images are handled equally well. [40] improves recent methods with the online hard example mining strategy.

Since humans are the ultimate receivers of images, the properties of the human visual system (HVS) should also be modeled in a data-driven manner. Perceptual error map is learned to guide quality prediction in [18]–[19], where DeepQA [18] is designed from FR-IQA methods, and BP-SQM [19] utilizes the U-Net to generate a similar map of the distorted image for reference. In HVS-Net [20], visual saliency and just noticeable difference (JND) [60] are taken into account to acquire the perceptually important features. Meantime, the rank loss is proposed to penalize the model when the order of its predicted quality scores is biased against that of the ground truth scores.

Some GAN-based methods have also been developed in the last few years [21], [22], [61]–[63]. H-IQA [21] and RAN4IQA [22] suppose that HVS unveils the mask of distortion and recreates a hallucinated scene without distortion in mind. In addition, AIGQA [62] proposes an active inference module based on the generative adversarial network (GAN) to predict the primary content. Since these IQA methods have achieved great improvement in synthetically distorted databases, a drawback exists when applied to authentic ones. The reference information is inevitably utilized in their training stage, which makes them limited in user generated content (UGC) due to the lack of reference images.

B. BIQA for Authentically Distorted Images

Most BIQA methods focus on synthetically distorted images, but relatively few approaches have been proposed to deal with the more challenging problem of authentic IQA. In recent years, based on multiple learning strategies, some methods are proposed to cope with this challenge. BLINDER [31] and DBCNN [30] pretrain a classification model on photographically generated classification databases such as ImageNet to acquire the quality-aware network parameters, which can help the regression task in IQA databases. MetalQA [32] adopts model-agnostic meta-learning (MAML) [64] to learn the prior knowledge among different synthetic distortions. However, due to the imbalance between the synthetic distortion and the authentic distortion, the learned knowledge cannot be generated effectively to the authentically distorted IQA databases, which is obvious in the experiment results. SFA-IQA [33] and HyperIQA [10] make IQA models understand the content diversity in authentic databases. The former utilizes semantic feature aggregation (SFA) to eliminate the impact of image content variation, and the latter utilizes a hyper-network architecture to evaluate the image quality adaptively according to the image content. In NAR-CNN [35], the authors propose a dual-path network to support IQA from a reference image with a similar scene but is not aligned. Considering the fact that an image receives divergent subjective scores from different human raters, PQR [37] and DeepRN [36] utilize the distribution of subjective scores to describe image quality. GraphIQA [38] develop Distortion Graph Representation (DGR) learning framework for BIQA, in which each distortion is represented as a graph and GraphIQA distinguishes distortion types by learning the contrast relationship between these different DGRs.

In this paper, as for the unanswered range effect in BIQA of UGC, we propose the novel REQA method, which has a more fine-grained ability to alleviate the range effect. Compared to the state-of-the-art methods, REQA can effectively tackle the prediction deviation in narrow ranges on the authentic databases, which contain images close to user generated content.

III. PROPOSED METHOD

To alleviate the range effect and improve the prediction performance in a narrow quality range, a novel BIQA method (named REQA) is proposed. As is illustrated in Fig. 2, REQA adopts a feedback hierarchy to realize coarse-to-fine quality assessment within multiple time steps. Here, a rank-and-gradient metric accomplishes coarse-grained assessment, and multi-stage pMOS refinements complete fine-grained assessment. In addition, REQA integrates image feature representations from two aspects, where the first is to process multi-scale distortion features through iteration and the second is the fusion of context features from Transformer Encoder [65].
Encoder for context understanding. Then we introduce two coarse-to-fine loss functions.

A. Feedback Network for Multi-scale Distortion Perception

Feedback Network (FN) is the core part of the proposed method, which controls the feedback process to perceive image degradation from multi-scale distortion features continually. The whole network consists of three elements, including feature extraction, time-domain iteration, and space-domain integration. The detail of this module is shown in Fig. 3.

1) Feature extraction: Multi-scale features contain diverse low-level information, which has been proved to be effective for IQA [13] [10]. Following [10], [38], [67], REQA adopts the ResNet-50 [66] as the backbone to acquire multi-scale features. Specifically, we remove global average pooling layers and fully connected layers of ResNet-50, and initialize corresponding network parameters using pretrained model in ImageNet. Finally, multi-scale features $f_{m1}$, $f_{m2}$, and $f_{m3}$ are extracted from $conv2_{-10}$, $conv3_{-12}$, $conv4_{-18}$ layers, respectively.

2) Time-domain iteration: Meantime, neurology [51] also proves that HVS can add details into distortion areas through feedback connection [52]. The property of HVS contributes to quality perception. Thus, multi-scale features combined with a feedback mechanism improve the fine-grained ability to quality perception. Thus, multi-scale features integrated and correlated with feedback connection [52]. The property of HVS contributes to quality perception. Thus, adding the context information to local feedback features contributes to understanding image quality comprehensively.

Compared to traditional CNN, Transformer [65] manages to capture long-range interactions thanks to its multi-head attention mechanism, that is, it is effective to obtain non-local features of an image. Meanwhile, CNN has some inherent inductive biases, such as translation invariant, scale invariant, and so on, which are not possessed by Transformer. These properties make CNN suitable for feature extraction. However, the layer occupies more computation resources compared to CNN. Therefore, we only apply one transformer layer on the top level of ResNet-50 to acquire the feedback context feature. The details of the Transformer Encoder (TE) are illustrated in Fig. 3. It consists of three elements: patch embedding, multi-head attention module, and multilayer perceptron.

1) Patch embedding: The original distorted image $I_i \in R^{H \times W \times 3}$ is processed by ResNet-50, and respective feature map $f_{m4} \in R^{H/4 \times W/4 \times C}$ is acquired by $conv5_{9}$. Here, $(H, W)$ is the resolution of $I_i$, and $C$ is the number of channels. Transformer encoder takes a one-dimensional vector as input. As for this, $f_{m4}$ is cut into $N$ patches firstly, where each patch $\mathcal{P}_{n,i} \in R^{P \times P \times C}$, $n \in [1, N]$ and $N = H \cdot W / (32^2 \cdot P^2)$. Then, $\mathcal{P}_{n,i}$ is flattened into a vector $\eta_{n,i} \in R^{1 \times P^2C}$, which is mapped into $\zeta_{n,i} \in R^{1 \times D}$ by a learnable matrix $W_{pe}$ as:

$$\zeta_{n,i} = \eta_{n,i} \cdot W_{pe}, W_{pe} \in R^{P^2C \times D} \quad (3)$$

where, in the experiment, we set $D = 224$. Moreover, to keep positional information, a standard learnable position embedding $\rho_{n,i} \in R^{1 \times D}$ is added into $\zeta_{n,i}$ to obtain the embedding $\xi_{n,i}$ of $\mathcal{P}_{n,i}$ as:

$$\xi_{n,i} = (\zeta_{n,i} + \rho_{n,i})^T \quad (4)$$

To realize the perception of the whole quality representation, we concatenate a trainable token $\mathbf{x}_{quality} \in R^{1 \times D}$ to the patch embeddings $\{\xi_{1,i}; \xi_{2,i}; \cdots ; \xi_{N,i}\}$, similar to BERT’s [69] and ViT’s [70] class token. Finally, the input matrix $X_i$ is defined as:

$$X_i = [\mathbf{x}_{quality}^T; \xi_{1,i}; \xi_{2,i}; \cdots ; \xi_{N,i}]^T \quad (5)$$

where $f_{id,t}$ is the input gate, $f_{gd,t}$ is the forget gate, $f_{od,t}$ is the output gate, $f_{cd,t}$ is the memory cell, $\sigma$ is the logistic sigmoid function, and $W_c$ is the weight matrix of conv block. In the proposed method, according to $d$, we set three conv blocks whose details are also shown in Fig. 4. The output $f_{od,t}$ is finally mapped into $v_{d,t} \in R^{1 \times 32}$.

B. Transformer Encoder for Context Understanding

To some extent, the perception of the local distortion can assess the degree of image quality degradation effectively. However, there are still some situations where local distortion cannot be quality-aware. For example, photographers are used to throwing the background out of focus to improve the visual effect of the foreground. In this case, when we only pay attention to the background and ignore its correlation with the foreground, it is easily regarded as a fuzzy distortion that affects the image quality. Thus, adding the context information to local feedback features contributes to understanding image quality comprehensively.
2) Multi-head attention: Following [65], self-attention map is acquired through a weighted sum over all values $V$ of the input matrix $X$. Each element in the weighting matrix $A$ are the pairwise correlation between two patch representations in $X$, which is calculated by the dot product of the respective query and key. The specific computational process is as:

$$[Q, K, V] = \mathcal{L}(X_i)W_{sa}, W_{sa} \in R^{D \times 3D_h}$$

$$A = \text{softmax}(QK^T), A \in R^{(N+1) \times (N+1)} \tag{6}$$

where $\mathcal{L}(\cdot)$ denotes layer normalization and $\mathcal{S}(\cdot)$ is self-attention operation.

Multi-head attention utilizes multiple self-attention operations to integrate global context relevance from local patch embeddings. Here, we set $D/D_h$ self attention as:

$$\mathcal{M}(X_i) = \mathcal{L}(\{S(X_1); S(X_2); \cdots; S(X_{D/D_h})\})W_{sa} + X_i \tag{7}$$

where $W_{sa} \in R^{D \times D}$ is the transition matrix.

3) Multilayer perceptron: We extract a 1D vector $\xi_{\text{quality}}$ from $\mathcal{M}(X_i)$ which corresponds to $x_{\text{quality}}$. Then MLP-I takes $\xi_{\text{quality}}$ as the input to obtain $\xi_{\text{quality}}$. After MLP-II processing $\xi_{\text{quality}}$, $v_m$ is computed. Finally, through feature concatenating and mapping, $F_\tau(I_t; \theta)$ is achieved as:

$$F_\tau(I_t; \theta) = MLP - III(v_m \oplus v_{1,t} \oplus v_{2,t} \oplus v_{3,t}) \tag{8}$$

where $\oplus$ represents concatenation operation.

### C. Coarse-to-Fine Loss Functions

The proposed method employs a multi-time strategy $t \in \{1, 2, \ldots, T\}$ where the feedback mechanism controls the progress of the BIQA task. In particular, it captures multi-scale quality-aware features, and then they are integrated and mapped into the quality score that is finally utilized by each time step to accomplish the assessment subtask. The whole assessment process is coarse-to-fine, containing the coarse-grained metric and fine-grained prediction.

1) Coarse-grained metric: As for coarse-grained metric, the proposed method redesigns the sampling strategy of training data. We randomly sample a mini-batch with $K$ 2-tuples $B = \{(I_k, s_k)\}_{k=1}^K$ from the current training set. Here, $I_k$ and $s_k$ are the training images and MOS labels, respectively. This randomly sampled mini-batch contains training samples from multiple quality scales. Coarse-grained metric aims to improve the perception of quality difference, so we group the training samples to form a micro-batch $B_m^n = \{(I_i, s_i)\}_{i=1}^n$ where a predicted image and four anchors are sampled from five different quality scales. Here, $m \in [1, M]$ where $M = \lfloor K/5 \rfloor$. Compared to the state-of-the-art ones [30] [33] [10] [37], the sampling strategy has two advantages as it can be adaptive to the metric task, and moreover, it can fix the small sampling problem in BIQA because of the combinatorial diversity.

The realization of coarse-grained metric is based on quality ranking and gradient keeping, which occupies the first-time step of feedback learning as the prior knowledge of fine-
grained prediction. Different from [20], [71], the proposed method puts forward a loss which can not only metric quality order but also keep the distance difference of the pairwise predicted scores consistent with one of respective ground truth scores. For \( I_i \) and \( I_j \), rank loss is first defined as:

\[
L_{rank}^{i,j}(t) = \text{max}(0, \frac{-s_i - s_j \cdot (F_{t-1}(I_i; \theta) - F_{t-1}(I_j; \theta))}{||s_i - s_j||_1 + \sigma})
\]  

(9)

where \( \theta \) is the network parameters, \( F_{t-1}(I_i; \theta) \) represents the quality prediction of image \( i \) in \( t = 1 \) and \( \sigma \) is a small stability term. In our experiment, we set \( \sigma = 0.0001 \). When the order of pairwise prediction sequence is consistent with the ground truth order, rank loss equal zero, otherwise, it is reduced to an absolute loss as:

\[
L_{rank}^{i,j}(t) = ||s_i - s_j||_1 \cdot ||F_{t-1}(I_i; \theta) - F_{t-1}(I_j; \theta)||_1
\]  

(10)

The rank loss keeps the order consistency, and gradient loss maintains the stability of pair-wised quality distance difference as:

\[
L_{gradient}^{i,j}(t) = ||s_i - s_j||_1 - ||F_{t-1}(I_i; \theta) - F_{t-1}(I_j; \theta)||_1
\]  

(11)

Overall, for a micro batch \( B^m_i \), coarse-grained loss is defined as:

\[
L_{t=1}^{Coarse} = \sum_{i=1}^{5} \sum_{j=i+1}^{5} (L_{rank}^{i,j}(i,j) + L_{gradient}^{i,j}(i,j))
\]  

(12)

2) Fine-grained prediction: Fine-grained prediction is inspired by the episodic curriculum learning [53], adopting an easy-to-hard strategy to realize \( t \in [2, T] \) prediction refinement. Specifically, We set a different threshold for each time step to compute the loss, and as the feedback information is processed continually, the threshold is reduced. Overall, the fine-grained loss is defined as:

\[
L_t^{Fine} = \text{max}(0, ||s_i - F_t(I_i; \theta)||_1 - l_t)
\]  

(13)

where \( L_t^{Fine} \) is the loss function in time step \( t \), and \( l_t \) represents the threshold in \( t \). The whole loss of the proposed method is calculated as:

\[
L = w_1L_{t=1}^{Coarse} + w_2L_{t=2}^{Fine} + w_3L_{t=3}^{Fine} + \cdots + w_T L_{t=T}^{Fine}
\]  

(14)

In our experiment, we set \( T = 4, l_2 = 5, l_3 = 2.5 \) and \( l_4 = 0 \). The values of \( w_1, w_2, w_3 \) and \( w_4 \) is dynamically adjusted, with the multi-time training task going on, to satisfy the curriculum learning principle. The set for \( w_1, w_2, w_3 \) and \( w_4 \) is in Section IV-A(3).

IV. Experiment

In this section, we first introduce the experimental protocols, including databases, criterion, and implementation details. Then we compare REQA with the state-of-the-art BIQA methods in terms of fine-grained prediction as well as traditional coarse-grained prediction performance, respectively. Next, we implement a series of ablation experiments to verify the contribution of different components of REQA. Finally, we also present some visualization samples acquired from the fine-grained distortion perception module to verify the effectiveness of the feedback hierarchy.

A. Experiment Protocols

1) Databases: Three authentically distorted IQA databases, including CLIVE [11], KonIQ-10k [41], and BID [42], are used to evaluate the performance of the proposed method. CLIVE contains 1162 images captured from diverse mobile
devices under real-world conditions. KonIQ consists of 10073 images, selected from 10 million YFCC100M entries. The sampling strategy of it concerns the authenticity of distortions, the diversity of content, and quality-related indicators. BID comprises 586 images with realistic blur distortions such as motion blur and defocusing blur, etc.

These databases are constructed based on quality ratings, i.e., bad, poor, fair, good, and excellent. Crowdsourcing strategy utilizes subjects to give multiple quality ratings to a single image. These ratings are then rescaled into [1, 100] to compensate for the biases of individual evaluations. Here, the higher score corresponds to higher quality. Through averaging them, MOS/DMOS is obtained as the quality label. In our experiment, because these three databases provide no original category labels, we divide the ground truth data according to Absolute Category Rating (ACR) [72] into five equal portions as Excellent (80 – 100), Good (60 – 80), Fair (40 – 60), Poor (20 – 40) and Bad (0 – 20) as for fine-grained prediction performance experiment.

2) Criterion: To evaluate the fine-grained ability to alleviate the range effect as well as the traditional coarse-grained ability, the Sperman Rank Order Correlation Coefficient (SROCC) and the Pearson Linear Correlation Coefficient (PLCC) are utilized as the evaluation metrics. SROCC is to measure the monotonicity between the ground truth data and the prediction scores. PLCC is to evaluate the linear correlation between these two. Given N images, the SROCC is defined as:

$$SROCC = 1 - \frac{6 \sum_{i=1}^{N} d_i^2}{N(N^2-1)}$$

(15)

where $d_i$ is the rank difference between MOS and pMOS of the $i$-th image. And PLCC is computed as:

$$PLCC = \frac{\sum_{i=1}^{N} (s_i - \bar{s})(\hat{s}_i - \hat{s})}{\sqrt{\sum_{i=1}^{N} (s_i - \bar{s})^2 \sum_{i=1}^{N} (\hat{s}_i - \hat{s})^2}}$$

(16)

where $s_i$ and $\hat{s}_i$ denote MOS and pMOS of the $i$-th image, and $\bar{s}$, $\hat{s}$ correspond to the mean of each.

3) Implement Details: We conducted both training and testing using Pytorch on an NVIDIA 2080 Ti GPU. In the proposed model, all the training images are deployed traditional data augment strategy, i.e. randomly crop the images to 224 × 224 pixel patches like [10], [32], [35]. The results are obtained from 20 train-test iterations. In each iteration, we randomly select 80% images for training, and the remaining 20% for testing, so there is no overlap between the training set and the test set. We train our model using Adam optimizer with weight decay 5e – 4 for 40 epochs. Learning rates for the backbone ResNet-50 and the other modules are first set to 2e – 5 and 2e – 4, respectively, and reduced by 10 in 10th epoch, 20th epoch, and 30th epoch, respectively.

In addition, the setting for $w_1$, $w_2$, $w_3$, and $w_4$ is consistent with the episodic curriculum learning [53]. First of all, we prioritize the assessment tasks corresponding to $w_1$, $w_2$, $w_3$, and $w_4$ in Eq.6 based on the easy-to-hard strategy. Specifically, the coarse-grained metric occupies the highest priority; Meanwhile, fine-grained loss with a bigger threshold possesses higher priority. Besides, we divide the training into four stages, with ten epochs for each stage. Finally, the setting is $w_1 = 0.25$, $w_2 = 0.5$, $w_3 = 0.25/3$, $w_4 = 0.25/3$ in stage 1, $w_2 = 0.5$, $w_3 = 0.25$, $w_4 = 0.25$ in stage 2, $w_2 = 0.25$, $w_3 = 0.5$, $w_4 = 0.25$ in stage 3 and $w_1 = 0.25$, $w_2 = 0.25$, $w_3 = 0.25$ and $w_4 = 0.5$ in stage 4.

B. Fine-grained Prediction Performance Experiment

In this subsection, we conduct the fine-grained prediction performance experiment on three datasets to evaluate the ability to eliminate the range effect. Compared methods includes HyperIQA [10], MetaIQA [32], and GraphIQA [38]. All the methods utilize the same division of the database to keep the consistency of test samples. Results from respective best models is evaluated in each category interval. For further examining the effectiveness of REQA, we also analyze the prediction bias and the number of outliers statistically.

1) Five quality ratings evaluation: As few methods focus on the fine-grained BIQA [47]. There exists no benchmark to explicitly evaluate fine-grained ability. Here, we divide three datasets (i.e., CLIVE, KonIQ-10k, and BID) into five fine-grained quality ratings following [72]: Excellent, Good, Fair, Poor, and Bad. As is shown in Tab. 1, REQA achieves outstanding performances in terms of all quality ratings. We can see that the range effect exists among almost all three methods: MetaIQA [32] performs well only on the Good and Fair quality ratings, which means if the input images are in the narrow range (i.e., Excellent, Fair, or Bad), the evaluation will degrade much; Also, HyperIQA [10] and GraphIQA [38] confront the same trouble that they both can not achieve ideal performance on the Excellent and Bad ratings. Compared with the three methods, REQA gets the best results in Poor, Fair, Good, and Excellent ratings. Especially in Poor rating, REQA achieves about 20% improvement both on SROCC and PLCC in contrast to GraphIQA. In the Excellent and Bad ratings, four methods all fail to obtain ideal results. According to statistic analysis, there are a few samples in these two ranges, which greatly increases the prediction difficulty as even an outlier leads to the obvious disturbance on SROCC. In this case, REQA still keeps competitive. In the Bad range, REQA achieves the best result on PLCC and the second-best result on SROCC. REQA also achieves the best in the Excellent rating.
Fig. 5: Visualization of predicted results. Green boxes focuses on the particular range and red boxes concerns the outliers.

|       | 0   | 1   | >2   |
|-------|-----|-----|------|
| REQA  | 1915| 442 | 7    |
| GraphIQA | 1796| 559 | 19   |
| HyperIQA | 1643| 710 | 21   |
| MetaIQA | 1692| 671 | 11   |

TABLE II: Statistics of prediction deviation in terms of Absolute Category Rating (ACR) scales.

|       | [0,2.5] | [2.5,5] | [5,7.5] | [7.5, 10] | >10 |
|-------|---------|---------|---------|-----------|-----|
| REQA  | 918     | 390     | 431     | 237       | 208 |
| GraphIQA | 771    | 390     | 473     | 237       | 265 |
| HyperIQA | 655    | 372     | 440     | 296       | 400 |
| MetaIQA | 404    | 483     | 461     | 268       | 213 |

TABLE III: Statistics of prediction deviation in terms of the prediction score.

2) Outlier and deviation analysis: To further verify the effectiveness of REQA, we make a statistic analysis between pMOS and the ground truth. As is shown in Fig. 5, we depict the experimental results on KonIQ-10K for direct visualization. On the whole, the linear fitting of the predicted results of REQA is more similar to the directly proportional function compared to MetaIQA [32] and HyperIQA [10]. Compared to GraphIQA [38], the distribution of its predicted results is more concentrated. This demonstrates that the results of REQA possess the best linear correlation with the ground truth data on a wide range. Moreover, focusing on a particular range (e.g., points within the green box), the results of REQA keep the same property. In addition, compared to the other methods, the predicted outliers (e.g., points within red boxes) of REQA is more close to the fitting curve. This means that REQA can alleviate the prediction deviation from a wide range effectively.

For further analysis, the prediction deviation of different quality ratings is quantified as illustrated in Tab. II. Here, we illustrate the statistic quantity of samples in terms of the quality ratings difference of Absolute Category Rating (ACR) [72] between pMOS and MOS. It is obvious that REQA is capable of limiting more images to its original rank scale, and meantime, it can reduce the wide range of prediction deviation as much as possible.

In addition, we also quantify the prediction deviation in terms of prediction score [0, 100], which is shown in Tab. III. Here, similar to Tab. II, we illustrate the number of the samples in terms of the prediction score differences of pMOS and MOS. Compared to the other methods, REQA achieves the best performance as there are 918 samples in [0, 2.5] and the most prediction biases concentrate in the first three ranges. This means that the predicted scores of REQA possess smaller fluctuation and furthermore demonstrates that the superiority of REQA for the fine-grained BIQA problem.

C. Coarse-grained Prediction Performance Experiment

In this subsection, we first conduct experiments on individual authentically distorted databases to verify the effectiveness...
of the proposed method in terms of the traditional coarse-grained ability [47], and then we make statistically significant test to validate the robustness of REQA. Lastly, we explore the generalization ability of the proposed method.

1) Single database evaluations: We compare the proposed REQA with 3 traditional methods and 11 DNN-based algorithms. The experimental results are exhibited in Tab. IV, where the top two SROCC and PLCC are marked in bold. All the results of the traditional methods are implemented where the top two SROCC and PLCC are marked in bold. The experiment is conducted by training on one database and testing on the full of another database.

| IQA methods | SROCC | PLCC |
|--------------|-------|------|
| BRISQUE [59] | 0.682 | 0.726 |
| ILMNIE [32] | 0.534 | 0.618 |
| HOSA [-]    | 0.633 | 0.711 |
| BIECON [73] | 0.595 | 0.618 |
| WaDIQAM [11]| 0.671 | 0.725 |
| SFA [-]      | 0.755 | 0.725 |
| DBCNN [-]   | 0.851 | 0.845 |
| SGDNet [67] | 0.851 | 0.872 |
| MetaloQA [37]| 0.802 | 0.825 |
| HyperIQA [10]| 0.859 | 0.869 |
| AIGQA [62]  | 0.751 | 0.781 |
| OLNNet [-]  | 0.849 | 0.858 |
| GraphIQA [38]| 0.845 | 0.870 |
| REQA        | 0.868 | 0.880 |

2) Generalization Ability Test: In order to explore the generalization ability of the proposed model, we run cross-database tests on authentically distorted databases CLIVE [11], KonIQ [41] and BID [42]. The best two results are highlighted in bold.

| IQA methods | SROCC | PLCC |
|--------------|-------|------|
| BRISQUE [59] | 0.682 | 0.726 |
| ILMNIE [32] | 0.534 | 0.618 |
| HOSA [-]    | 0.633 | 0.711 |
| BIECON [73] | 0.595 | 0.618 |
| WaDIQAM [11]| 0.671 | 0.725 |
| SFA [-]      | 0.755 | 0.725 |
| DBCNN [-]   | 0.851 | 0.845 |
| SGDNet [67] | 0.851 | 0.872 |
| MetaloQA [37]| 0.802 | 0.825 |
| HyperIQA [10]| 0.859 | 0.869 |
| AIGQA [62]  | 0.751 | 0.781 |
| OLNNet [-]  | 0.849 | 0.858 |
| GraphIQA [38]| 0.845 | 0.870 |
| REQA        | 0.868 | 0.880 |

3) Ablation Study

In this subsection, we make ablation experiments to verify the contribution of key constituent parts of REQA. We conduct ablation experiments on KonIQ [41]. The training and testing protocols are the same as above.

| SROCC | Excellent | Good | Fair | Poor | Bad |
|-------|-----------|------|------|------|-----|
| w/o FN+TE | 0.115 | 0.651 | 0.595 | 0.640 | 0.332 |
| w/o TE  | 0.238 | 0.726 | 0.642 | 0.703 | 0.560 |
| w/o FN  | 0.153 | 0.691 | 0.630 | 0.668 | 0.346 |
| w/ t=2  | 0.217 | 0.722 | 0.649 | 0.689 | 0.463 |
| w/ t=3  | 0.228 | 0.740 | 0.662 | 0.701 | 0.485 |
| w/ t=5  | 0.252 | 0.752 | 0.664 | 0.720 | 0.573 |
| REQA   | 0.252 | 0.751 | 0.665 | 0.722 | 0.571 |

| PLCC | Excellent | Good | Fair | Poor | Bad |
|------|-----------|------|------|------|-----|
| w/o FN+TE | 0.340 | 0.693 | 0.596 | 0.684 | 0.776 |
| w/o TE  | 0.425 | 0.732 | 0.643 | 0.754 | 0.838 |
| w/o FN  | 0.349 | 0.715 | 0.633 | 0.734 | 0.796 |
| w/ t=2  | 0.372 | 0.728 | 0.647 | 0.753 | 0.830 |
| w/ t=3  | 0.405 | 0.740 | 0.665 | 0.762 | 0.846 |
| w/ t=5  | 0.436 | 0.745 | 0.659 | 0.769 | 0.853 |
| REQA   | 0.435 | 0.746 | 0.660 | 0.768 | 0.851 |
TABLE VII: Ablation Experiments about the effectiveness of coarse-grained loss and fine-grained loss. The comparison is conducted based on the fine-grained prediction performance.

|                   | Excellent | Good | Fair | Poor | Bad |
|-------------------|-----------|------|------|------|-----|
| w/o $L_{Coarse}$ | 0.233     | 0.728| 0.626| 0.686| 0.558|
| w/o $L_{Fine}$   | 0.213     | 0.716| 0.609| 0.659| 0.536|
| REQA              | 0.252     | 0.751| 0.665| 0.722| 0.571|
| PLCC              | Excellent | Good | Fair | Poor | Bad |
| w/o $L_{Coarse}$ | 0.398     | 0.725| 0.644| 0.743| 0.827|
| w/o $L_{Fine}$   | 0.365     | 0.704| 0.629| 0.725| 0.807|
| REQA              | 0.435     | 0.746| 0.660| 0.768| 0.851|

The mechanism is conducive to different quality ratings, especially for Excellent and Bad. Moreover, to further analyze the effectiveness of the feedback mechanism in FN module, we change the number of time steps (i.e., w/ t=2;3;5). We can see that with the increasing of number of time steps, the performance gains in terms of Excellent and Bad quality ratings are more than other quality ratings, which validates that the feedback mechanism can gradually refine the fine-grained prediction results. Besides, w/ t=4 reaches the optimal point between performance and computation efficiency. Our proposed model (REQA) achieves the best results, which illustrates that modules interact with each other to get a positive gain.

Then we analyze the effectiveness of $L_{Coarse}$ and $L_{Fine}$, which is shown in Tab. VII. We replace the $L_{Coarse}$ (w/o $L_{Coarse}$) and $L_{Fine}$ (w/o $L_{Fine}$) with $L_1$ loss, respectively. According to the fine-grained performance comparisons, we can conclude that coarse-to-fine loss functions, especially for $L_{Fine}$, benefit from different quality ratings.

Overall, we can make the subsequent conclusions. First of all, different modules proposed in this paper can improve the performance of REQA. TE enlarges the receptive field to acquire the context features. FN can obtain a more quality-aware perception of multi-scale features while completing the task iterations. $L_{Fine}$ makes REQA have a prediction accuracy and with the addition of $L_{Coarse}$, fine-grained prediction performance gets promoted. All of these are responsible for the outstanding experimental results of REQA.

E. Visualization of Fine-grained Distortion Features

In this subsection, we visualize the fine-grained distortion information in the form of the heat map [77] to further examine the performance of the feedback hierarchy. We show each heat map corresponding to the output of $d^{th}$ feedback block in $t^{th}$ time step in Fig. 6. Here, the displayed images are from KonIQ [41]. As we can see from the longitudinal comparison, different feedback blocks possess different receptive fields. For example, by comparing Fig. 6 (a), (e), and (i), it can be concluded that the third feedback block perceives larger areas where the textures are much clearer. This is attributed to different scales of inputs of three feedback blocks. Furthermore, by horizontal comparison, it can be concluded that feedback blocks can obtain more fine-grained feature maps, which is consistent with our standpoint. For example, through making comparisons of Fig. 6 (a)-(d), we can see that the heat map in $t = 4$ has more heat sensitive areas (e.g., points within green boxes). In addition, we obtain the pMOS of the image in each time step as 79.01 in $t = 1$, 82.02 in $t = 2$, 83.16 in $t = 3$, and 82.13 in $t = 4$. The MOS of it is 84.91. This means that with the feedback going on, REQA can make a more accurate prediction. Overall, we can conclude that the feedback hierarchy can achieve better and better features as iteration goes on. These features are quality-aware enough to ensure REQA obtain more fine-grained prediction result.

V. Conclusion

In this paper, to our best knowledge, we take the first attempt to develop a fine-grained blind image quality assessment method to alleviate the range effect. Concretely, we first propose the coarse-to-fine method with a strong fine-grained prediction ability. Benefiting from Feedback Network (FN) and Transformer Encoder (TE), the proposed method can perceive the multi-scale distortion information and global context information, which makes the model quality-aware for fine-grained distortions. Furthermore, by integrating coarse-grained metric and fine-grained losses into the feedback hierarchy to process these features, the proposed method achieves outstanding coarse-grained and fine-grained prediction performance, as is demonstrated by a series of experimental results. In future work, we will try to develop a metric to quantitatively evaluate the range effect via statistical methods, which can also be applied to evaluate the fine-grained prediction ability of existing IQA methods.

REFERENCES

[1] Z. Pan, F. Yuan, J. Lei, Y. Fang, X. Shao, and S. Kwong, “Vcrnet: Visual compensation restoration network for no-reference image quality assessment,” IEEE TIP, vol. 31, pp. 1613–1627, 2022.
