IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 7 September 2022, accepted 19 September 2022, date of publication 26 September 2022, date of current version 3 October 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3209810

RESEARCH ARTICLE

Convolved Quality Transformer: Image Quality Assessment via Long-Range Interaction Between Local Perception

HEESEOK OH[1,*], JINWOO KIM[2,*], TAEWAN KIM[3], AND SANGHOON LEE[2], (Senior Member, IEEE)

[1]Department of Applied AI, Hansung University, Seoul 02876, South Korea
[2]Department of Electrical and Electronic Engineering, Yonsei University, Seoul 03722, South Korea
[3]Data Science Major, Dongduk Women's University, Seoul 02748, South Korea

Corresponding authors: Taewan Kim (kimtwan21@dongduk.ac.kr) and Sanghoon Lee (slee@yonsei.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) Grant funded by the Korea Government (MSIT) under Grant NRF-2020R1G1A1100674.

*Heeseok Oh and Jinwoo Kim are co-first authors.

ABSTRACT A hybrid architecture composed of a convolutional neural network (CNN) and a Transformer is the new trend in realizing various vision tasks while pushing the limits of learning representation. From the perspective of mechanisms of CNN and Transformer, a functional combination of them is suitable for the image quality assessment (IQA) since which requires leveraging both local distortion perception and global quality aggregation, however, there has been scarce study employing such an approach. This paper presents an end-to-end CNN-Transformer hybrid model for full-reference IQA named convolved quality transformer (CQT). The CQT is inspired by the human's perceptual characteristics and is designed to unify the advantages of both CNN and Transformer for evaluating quality score. In CQT, convolutional layers specialize in local distortion feature extraction whereas Transformer aggregates them to estimate holistic quality via long-range interaction between them. Such a series of processes is repeated on multi-scale feature maps to capture quality representation sensitively. To verify submodules in CQT perform their roles properly, we in-depth analyze the interaction between local distortions inferring global quality with attention visualization. Finally, the perceptually pooled information from stage-wise feature embeddings derives the final quality level. The experimental results demonstrate that the proposed model achieves superior performance in comparison to previous data-driven approaches, and which is even well-generalized over standard datasets.

INDEX TERMS Full-reference image quality assessment, human visual system, CNN-transformer hybrid model.

I. INTRODUCTION

THE goal of image quality assessment (IQA) is to predict an objective score of the given image equivalent to that perceived by humans. The IQA tasks have continued to receive attention since objective image quality is able to play a role as a quantitative criterion without human judgments while developing an end-display product, producing visual content, and improving the associated service. In several practical scenarios, visual information suffers a wide variety of distortions during acquisition, compression, transmission, rendering, etc [1]. An objective approach for an evaluation of image quality is to quantify the visible differences between the reference and distorted images [2]. In this context, IQA methods are generally categorized into three approaches in accordance with the presence or absence of a reference image (pristine image in general), full-reference IQA (FR-IQA), reduced-reference IQA (RR-IQA), and no-reference IQA (NR-IQA) [3]. Although RR- and NR-IQA methods are preferred for practical usage as an evaluation criterion, owing to the benefits of the relative information delivered by a pair-wise comparison between the utilized reference and

The associate editor coordinating the review of this manuscript and approving it for publication was Yizhang Jiang.

target distorted images, the FR-IQA method guarantees to achieve superior predictive power than them [4].

The quality prediction accuracy is determined by the consistency degree with the human's opinion. Almost of IQA dataset provides an image quality ground-truth as a mean opinion score (MOS) which is an averaged value over subjective scores rated by several participants. That is, in order to achieve the construction of a FR-IQA metric/model highly correlated with MOS, it is an ultimate factor that how the perceptual characteristics based on the human visual system (HVS) can be well considered to evaluate the quality score [5]. Hereby, various previous works attempted to design IQA metrics involved in the lower- and higher-level HVS [2], [6], [7], [8], [9].

Nevertheless, because of the unfathomable complex mechanism of HVS, it is difficult to model high-performance IQA evaluators by reflecting some of the fragmentary factors. For this reason, in the past decade, data-driven approaches based on CNN (convolutional neural network) rapidly emerged [10], [11], [12], [13], [14]. The key idea underlying CNN-based IQA is local connectivity to extract meaningful representations of local distortions. By stacking multiple convolutional layers, the effective receptive fields may enlarge to capture global quality-related characteristics [15]. To cope with the weakness caused by the inductive bias underlying CNN, in more recent years, Transformers starts unleashing the power in entire computer vision fields [16], [17], [18], and several Transformer-based IQA models were also introduced [19], [20], [21], [22], [23]. Such existing works for Transformer-based IQA methods were all early studies that just applied Transformer architecture to their quality predictors without any consideration of the perceptual characteristics of IQA.

In this paper, we introduce a learning-based FR-IQA model, *convolved quality transformer (CQT)*, which deploys a hybrid architecture composed of CNNs and Transformers to achieve reliable performance by reflecting visual characteristics regarding that a human perceives degradation in an image over both local and global regions together to determine quality level. As shown in Fig. 1, some distortions are highlighted and some others are less noticeable even though the same level distortions are uniformly distributed over all spatial regions (i.e., globally added Gaussian noise as seen in the absolute difference). In other words, we easily perceive the distortions in homogeneous regions (e.g., sky and road), but those in textural regions (e.g., trees and gardens) are less sensitive. After observing such locally regional distortion, human aggregates all perceived information and interpret it to a higher level by considering their global relationship to determine the holistic image quality [4], [24].

In terms of designing the proposed CQT, such observation inspires that the representations of local distortions are differently abstracted following the corresponding region characteristics in spite of the equal distortion level, and it is well-known that the convolutional layer is powerful to capture the local features to understand the overall image

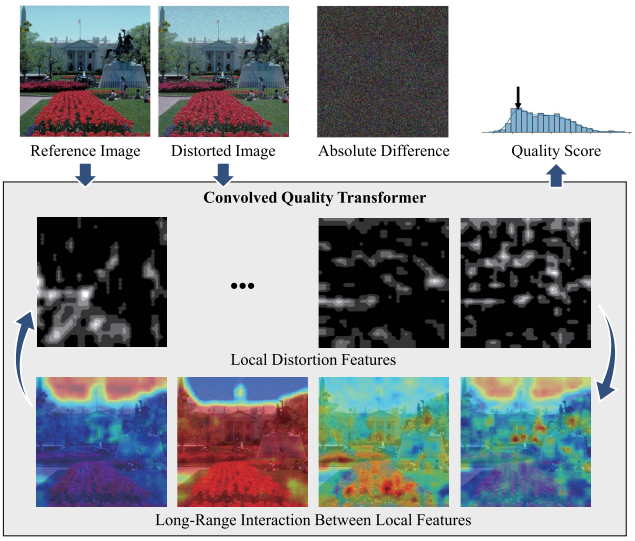


FIGURE 1. Overview of the proposed image quality evaluator. The reference and distorted images are fed into the model to extract local distortion features. The model estimates the long-range relationship between every local information and predicts the degree of image quality.

with ignoring the contextual relation of them that are outside of the receptive field [25], [26]. After that, the extracted local features have to be aggregated into the global features with capturing their long-range dependencies. However, it is difficult for local information to be interacted with and merged by using only the convolutional layers which have a small field of view to capture global quality, thus previous IQA approaches were explicitly pool information by taking an average [4], [27]. Whereas, ViT is designed to exhaustively encode long-range dependencies between local regional information over an image through multi-head self-attention (MHSA) [16]. In CQT, MHSAs are employed to empower the model with the ability to long-range interact between local quality perceptions, and convolutional layers are also embedded in between to extract regional distortion features affecting the degree of perceived quality, as depicted in Fig. 1.

The proposed CQT leverages a convolutional projection into the Transformer block to maximize the advantage of utilizing both CNN and Transformer which maintains local distortion perception and long-range interaction between them. Moreover, our model consists of the image encoders, CQT blocks, and a prediction head utilizing perceptual poolings, which is designed stage-wise manner to efficiently build multi-scale feature maps regarding addressing coarse and fine-grained distortions. The main contribution of this work is three-fold: (1) We newly introduce CQT, a novel learning-based FR-IQA model composed of the hybrid structure which takes advantage of both CNNs and Transformers. (2) We demonstrate by visualization that the proposed model is possible to extract local distortion features effectively and to understand their long-range interactions over spatial regions similar to HVS characteristics. (3) We present comprehensive

ablations studies of the model architecture and verify the superiority of CQT outperforming previous works on LIVE [28], CSIQ [29], TID2013 [30], TID2008 [31] IQA datasets.

II. RELATED WORK

In order to clarify the status and limitations of the IQA field, we state the previous works in three categories–conventional metrics reflecting HVS, CNN-based data-driven approaches, and Transformer-based learning methods.

A. FUNDAMENTAL APPROACHES FOR HVS-BASED IQA

The conventional metrics on IQA focused on imitating HVS as a closed-form weight function and applying it to estimate the pixel-wise error between reference and distorted images. SSIM [2] is a de-facto standard index in image processing fields which reflects that HVS is sensitive to structural information. Tsai and Liu [6] demonstrated that the non-uniform resolving power of the retina (i.e., foveation) weighting guides determining the overall quality of an image. FSIM [7] showed that HVS understands an image via local perception, which is highly correlated with phase congruency. Zhang *et al*. [8] proposed the VSI index which computes local distortion quality based on visual saliency and pools them. Laparra *et al*. [32] assumed that image quality is dominated by local luminance error according to the early visual pathway (area V1), and applied Laplacian pyramid decomposition into preprocessing. Xue *et al*. [33] estimated the local quality of each small patch in the distorted image based on gradient magnitude similarity and pooled the values to derive the final score.

Several HVS-based IQA methods have been proposed, but there is no clear definition of structural distortion in a perceptual meaning, and most methods were limited to their assumed functional forms regarding the interaction between local distortions. Despite such limitations, previous HVS-related IQA studies share a common two-step framework–local distortion extraction at first and then computing the global quality. This fundamental strategy inspired us to design CQT model which implicitly integrates both learning local distortion features and understanding their dependency to determine the holistic quality score in a perceptual manner.

B. CNN-BASED IQA METHODS

In the past decade, CNN-based IQA models were actively studied. The first CNN-based method was used for NR-IQA. Kang *et al*. [27] employed a patch-wise approach where divided and locally normalized patches were fed into a shallow CNN to craft features and the pooled value was supervised by MOS. Liang *et al*. [10] proposed a dual-path FR-IQA model using locally normalized patches. They employed weight-sharing CNNs for non-aligned reference and distorted image pairs, and the concatenated learned vectors were regressed onto subjective scores. Bosse *et al*. [34] subdivided images into patches and estimated local patch-wise qualities. And then to estimate image-wise quality by aggregating them, similar to DeepQA [35], the visual weight for each patch was learned during the model training.

However, such local patch-based learning approaches had a weakness in estimating a global score since the aggregating over the obtained local information was performed without any prior relation to human perception. To handle this issue, more recent studies explored the semantic feature-level IQA methods. Gao *et al*. [11] studied how mid-level representations of pre-trained VGGNet can be used to determine image quality. The local similarity between the extracted feature maps from reference and distorted images was calculated, and the global picture-quality scores were pooled. LPIPS [12] showed that the calibrated ℓ_2 distance between reference and distorted images on the feature space learned by a deep network represents perceptual similarity measure, i.e, image quality. PieAPP [36] utilized a pairwise learning framework for IQA which learns the preference of one distorted image over the other based on an estimated perceptual error from the extracted features. DISTS [14] considered both structural and textural similarities. This metric was explicitly designed to tolerate texture resampling where the score is measured in a SSIM-like way between feature embeddings mapped by the pre-trained CNN.

C. TRANSFORMER-BASED IQA METHODS

Transformer network is being applied in a wide vision field, hereby various Transformer-based IQA techniques have been also proposed. You and Korhonen's work [20] was the first attempt to utilize Transformer for NR-IQA. The extracted features were linearly projected and an extra learnable embedding of the Transformer encoder was regressed onto MOS. To imitate human behavior, Zhu *et al*. [37] attached a saliency detector, and the estimated region of interest on feature level was served to MHSA as a query for NR-IQA. Ke *et al*. [38] introduced a multi-scale embedding approach including hash-based 2D spatial embedding and a scale embedding strategy to handle various resolutions and aspect ratios for the generalized NR-IQA. Cheon *et al*. [19] expand Transformer-based long-range understanding to FR-IQA. The semantic feature difference was extracted by using a pre-trained CNN classifier backbone and was passed to Transformer to predict the final score. Jiang *et al*. [21] calculated the non-linear residuals between distorted and reference features captured from the CNN backbone, and which were fed into Transformer layers for quality evaluation. Chubarau and Clark [22] employed the probabilistic sampler to extract multi-scale patches implying representative quality difference, and then the MHSA modules encoded long-range dependencies between those embedded patches. Keshari *et al*. [23] adopted the bagging ensemble method to cope with multi-scale issues in IQA. The Transformer-based models were trained on different image scales, and the inferred quality scores were averaged. However, the previous studies have been limited to naïve approaches to applying Transformer to the IQA field as the regressor ignoring the specialized characteristics to determine perceptual quality.

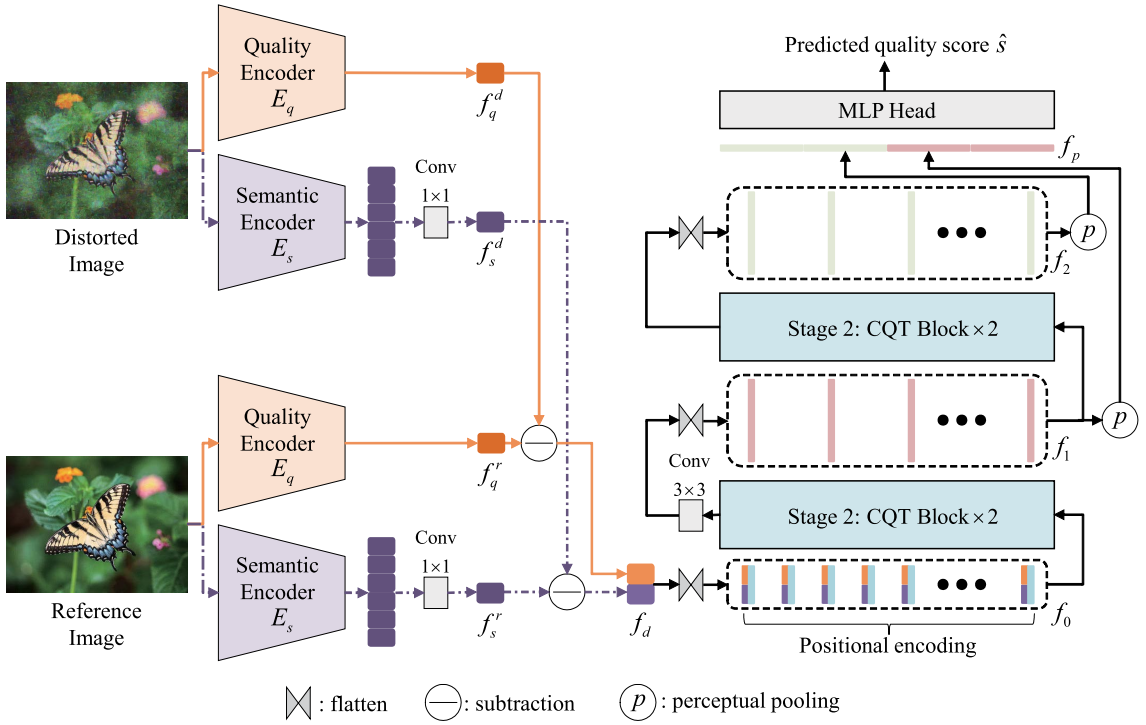


FIGURE 2. The proposed CQT model architecture for FR-IQA.

III. PROPOSED CQT MODEL FOR FR-IQA

The proposed IQA model has a hybrid structure that aims for taking advantage of Transformers to extract long-range dependency, and that of CNNs to capture a quality representation of the local regions. The proposed model architecture is depicted in Fig. 2 which consists of three substructures–image encoding CNNs, CQT blocks, and a score predictive MLP (multi-layer perceptron) head. Both distorted and reference images are encoded into the lower-dimensional representations by the shared CNNs, and the subtracted features of them are fed into the CQT to reason an overall quality degradation. Here, the CQT block employs the MHSA module and several convolution layers which is for an effective understanding of the quality difference between the distorted and reference images in terms of both holistic and local perception. After capturing short- and long-range dependencies over the multi-scale quality feature difference through multi-stage CQT blocks, those are pooled as the final features where a perceptual pooling function is designed to reflect the human visual characteristics. Finally, the MLP head predicts the degree of quality as an objective score by using them.

A. IMAGE ENCODER

In order to reduce the dimension, both distorted and reference images are encoded by CNNs sharing the parameters. In previous approaches, an image is generally split into the non-overlapped patches for the construction of a token sequence when a Transformer-based model is utilized [16], [39], [40], [41], [42], [43]. Such a patch-based way has a problem that it is difficult to achieve extracting local perceptual features. To cope with this, recent approaches adopted early convolution for information abstraction [44], [45], [46], [47], [48]. The encoded spatial information is treated as a token sequence on the feature level, and it is passed to the Transformer for understanding its long-term relationship, thus the CNN plays the role of the kind of learnable image downsampler.

In the proposed model, the input images are encoded through the dual path. Here, two encoders are employed including the quality encoder E_q and the semantic encoder E_s. First of all, the role of both encoders is the same in that they reduce the dimension and generate a token sequence, however, their fundamental purposes are distinct.

1) QUALITY ENCODER E_q

The quality encoder E_q is designed to extract the representative distortion features, and which encodes the pristine and distorted images $\left(I^r, I^d \in \mathbb{R}^{h\times w\times 3}\right)$ to feature maps f_q^r and f_q^d having the reduced spatial size.

$$\begin{aligned} f_q^r &= E_q\left(I^r;\theta_q\right), \\ f_q^d &= E_q\left(I^d;\theta_q\right), \end{aligned} \tag{1}$$

where θ_q indicates the parameters of E_q. Here, E_q is a trainable network and its architectural design is inspired by the encoder of VQGAN [49] which effectively reduces the spatial dimension of input images, from 256 × 256 to 32 × 32.

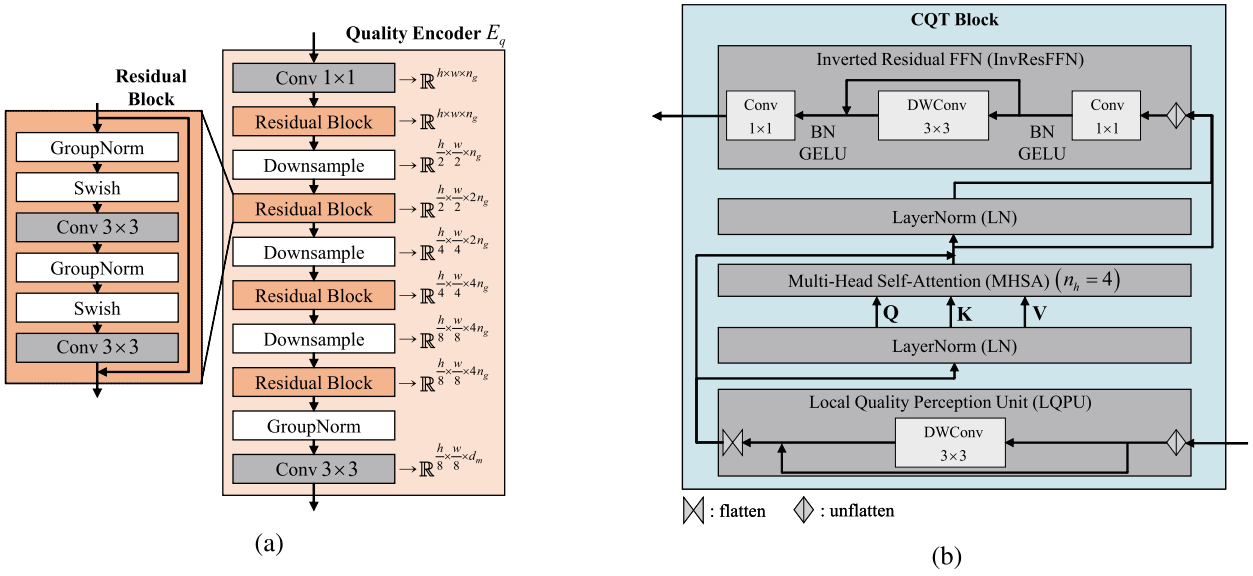


FIGURE 3. The detailed layer structure of the proposed model. (a) the quality encoder E_q. (b) CQT block.

The detailed architecture of E_q is depicted in Fig. 3 (a). At first, an image channel is expanded to $n_g = 64$ by convolution, and the final extracted feature maps f_q^r and f_q^d have the shape $\mathbb{R}^{\frac{h}{8}\times\frac{w}{8}\times d_m}$, where the embedding dimension d_m is set to 128. After encoding both reference and distorted images by the weight-sharing CNNs, the degraded quality feature is obtained by subtracting two features, i.e., $f_q^r - f_q^d$. This encoding path is mainly to capture the degree of local quality degradation, thus E_q is trained to resolve distortion sensitively.

2) SEMANTIC ENCODER E_S

It is well known that an unnaturally corrupted semantic region in an image attracts the attention of a human [50], and which implies that the semantic distance is also important information to determine the image quality. Human judgment of similarity depends on high-order image structure, thus several previous works utilized the semantic features to predict image quality and to reflect perceptual characteristics on image restoration tasks (e.g., perceptual loss) [12], [19], [51], [52]. Thus, we employed the pre-trained Inception-ResNet-V2 [53] on ImageNet dataset as the semantic encoder E_s. Differ from E_q, the parameters θ_s of E_s are frozen and not updated during the model training, and which extracts the semantic feature maps f_s^r and f_s^d from both reference and distorted images.

$$\begin{aligned} f_s^r &= E_s\left(I^r;\theta_s\right), \\ f_s^d &= E_s\left(I^d;\theta_s\right), \end{aligned} \tag{2}$$

The total six feature maps are obtained by the intermediate layers in E_s (i.e., *mixed_5b*, *block35_2*, *block35_4*, *block35_6*, *block35_8*, *block35_10* layers) where the same setting as used in the model proposed by Cheon *et al.* [19]. The feature maps are concatenated, and whose representation along the channel dimension is compressed by applying 1×1 convolution. As a result, the semantic feature maps f_s^r and f_s^d have the same shape $\mathbb{R}^{\frac{h}{8}\times\frac{w}{8}\times d_m}$ as the quality feature maps f_q^r and f_q^d. Similar to the quality encoding path, the final output of the semantic path is the difference between the features encoded from reference and distorted images, i.e., $f_s^r - f_s^d$.

Note that in our structure, the trainable quality encoder E_q plays the dominant role in capture the quality degradation, and the semantic encoder E_s complements it while the pre-trained network in previous works has been generally utilized as the main backbone to extract the features [12], [19].

B. CQT BLOCK

Several studies incorporating convolutions into the vision Transformer models are emerging [45], [54], [55], and there are infinite possibilities enabling the combination of CNN and Transformer regarding model architecture. The proposed CQT is also a type of hybrid structure applying convolution to Transformer whose main purpose is understanding long-range interaction between locally degraded quality by self-attention to estimate global quality. As shown in Fig. 2, the encoded feature differences $f_q^r - f_q^d$ and $f_s^r - f_s^d$ are concatenated as

$$f_d = \left[f_q^r - f_q^d, f_s^r - f_s^d\right], \tag{3}$$

where $f_d \in \mathbb{R}^{\frac{h}{8}\times\frac{w}{8}\times 2d_m}$. f_d is spatially flatten for being token sequence having $N = \frac{h}{8} \times \frac{w}{8}$ length, and then an additional learnable embedding $p_d \in \mathbb{R}^{N\times 2d_m}$ is added to this as usually used in the Transformer based models for maintaining the

positional information as

$$f_0 = \left[f_{d_0} + p_{d_0}, f_{d_1} + p_{d_1}, \ldots, f_{d_N} + p_{d_N}\right], \tag{4}$$

and then CQT identify the relationship between quality information of each token contained in f_0.

Our CQT block is mainly inspired by the work of Guo *et al* [48], and which consists of local quality perception unit (LQPU), MHSA module, and an inverted residual feed-forward network (InvResFFN). The overall architecture of CQT block is presented in Fig. 3 (b). It is well known that convolution and MHSA compensate for drawbacks of each other, and the convolutional layer in an early stage is better at processing local patterns [54]. In this context, the LQPU attempts to understand the quality relation and the structural information contained in a local spatial region (i.e., unflattened token sequence). The LQPU deploys the depth-wise convolution and retains the feature dimension. The output feature of the LQPU is normalized and fed into the MHSA after linear projections to query q_l, key k_l, and value k_l at l^{th} block. The MHSA aims for long-range interaction between the locally perceived quality degradation whose the number of multi-head n_h is set to 4. The shortcut from LQPU is inserted to promote gradient propagation, and the output of the MHSA is inputted to InvResFFN where the expanding ratio is set to 4. The calculation of the CQT block is formulated as:

$$\begin{aligned} x_0 &= f_0, \\ x_l' &= LN\left(LQPU\left(x_{l-1}\right)\right), \\ q_l &= k_l = v_l = x_l', \\ x_l'' &= MHSA\left(q_l, k_l, v_l\right) + LQPU\left(x_{l-1}\right), \\ x_l &= InvResFFN\left(LN\left(x_l''\right) + x_l''\right). \end{aligned} \tag{5}$$

Note that we adopt the hierarchical (i.e., multi-stage) Transformers structure which deals with the convolutional features in the token embeddings with attention projection since it is shown that such a multi-scale setting leads the improved performance over several image reasoning tasks [40], [45]. As depicted in Fig. 2, The proposed model consists of a total of two stages, and each stage has two CQT blocks, respectively. The output feature of the first stage is projected to token embeddings $f_1 \in \mathbb{R}^{\frac{h}{16} \times \frac{w}{16} \times 4d_m}$ which is formed by 3×3 convolution. f_1 is spatially reduced in half compared to f_0 which is fed into the second stage and the output token embeddings $f_2 \in \mathbb{R}^{\frac{h}{32} \times \frac{w}{32} \times 8d_m}$ can be obtained, and then both f_1 and f_2 is used to estimate the final degree of perceptual image quality.

C. MLP PREDICTION HEAD

Through the MLP prediction head at the end of the model, final objective image quality score is estimated based on the extracted feature embeddings f_1 and f_2. Our MLP prediction head consists of two linear layers with having $12d_m$ latent variables. Here, f_1 and f_2 are indirectly fed into the MLP prediction head, we employed two types of pooling strategies including the global average pooling and p-percentile pooling. It is well-known that taking an average represents the global image quality in holistic manner [27], [56]. In addition, the most severe distortion of visual content has a dominant effect on the overall perceived quality [57], thus we capture such spatially localized severe distortions by averaging the upper p^{th} percentiles from the feature value distribution in order to reflect human's visual characteristics [24]:

$$p_m\left(f\right) = \frac{1}{N_t} \sum_{i=1}^{N_t} f_{i,j}^h, \tag{6}$$

$$p_p\left(f\right) = \frac{1}{N_t^p} \sum_{i > n^{p+}} f_{i,j}^h, \tag{7}$$

where i and j are spatial (i.e., token) and channel (i.e., embedding dimension) indices of the input feature f, and n^{p+} represents the upper p^{th} percentiles in the histogram of feature values f^h. N_t and N_t^p indicate the total number of feature values and the number of p-percentile local qualities (i.e., $N_t^p = N_t \cdot p/100$), respectively, and we set $p = 5$.

As presented in Fig. 2, the feature embeddings at different stages f_1 and f_2 are separately pooled and concatenated to keep the multi-scale characteristics of image distortion. Thus, four pooled feature vectors $p_m\left(f_1\right) \in \mathbb{R}^{4d_m}$, $p_p\left(f_1\right) \in \mathbb{R}^{4d_m}$, $p_m\left(f_2\right) \in \mathbb{R}^{8d_m}$ and $p_p\left(f_2\right) \in \mathbb{R}^{8d_m}$ can be obtained, and which are concatenated together to form the final feature $f_p \in \mathbb{R}^{24d_m}$. The MLP head predicts the final image quality score $\hat{s}$ from f_p.

IV. EXPERIMENTS

A. DATASETS

In order to verify the predictive power of the proposed method, five public IQA databases were utilized in our experiments.

- KADID-10k [58]: contains 81 reference images and 10,125 distorted images using 25 fine-grained distortion types including blurs, color distortions, compression, noise, brightness changes, sharpening, contrast change, and the other spatial distortions. A total of 30.4k subjective ratings were collected by crowdsourcing.
- CSIQ [29]: contains 30 references with 866 distorted images using six types having five levels of distortions (i.e., JPEG, JPEG2000, contrast decrements, Gaussian noise, and Gaussian blur) where whose 5k subjective scores are from the lab environment.
- LIVE [28]: contains 29 pristine images and whose 779 distorted versions by applying five synthetic distortion types (i.e., JPEG, JPEG2000, white noise, Gaussian blur, and fast fading). To construct MOS, a total of 25k ratings are collected under the lab environment.
- TID2013 [30]: contains 25 pristine images and whose 3k distorted versions using 25 distortion types where each having five levels of degradations, and the MOSs were constructed by 524k subjective ratings under the lab environment.
- TID2008 [31]: contains 25 references and their 1700 distorted images using 17 types of distortions having

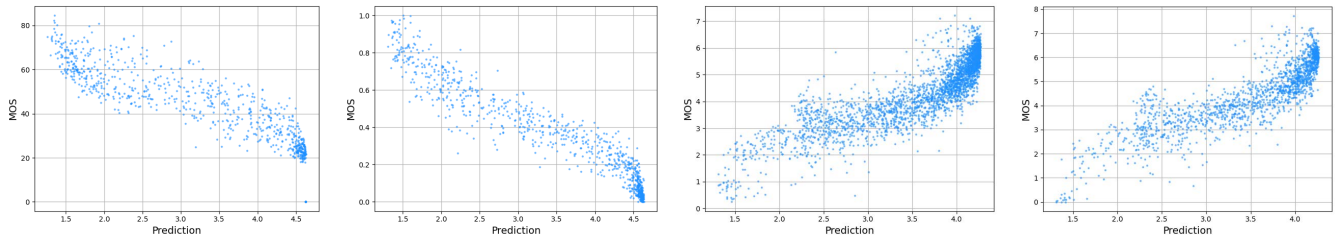


FIGURE 4. Scatter plots of MOS against the predicted quality score of the CQT on three benchmark datasets, (a) LIVE, (b) CSIQ, (c) TID2013, and (d) TID2008.

four levels. The subjective test was performed by 838 observers in the lab environment.

These datasets are de-facto standards for FR-IQA, and we determined to use KADID-10k as a training dataset in this study since its scale is larger compared to the others, and ratings were collected from crowdsourcing.

B. IMPLEMENTATION DETAILS

Each reference and distorted image pair was randomly cropped as 256 × 256 size, horizontally and vertically flipped, and rotated in the training step. The MOS value of the original image was employed as a supervisor for each cropped patch [19], [23]. All of the images were normalized with 0.5 mean and 0.5 standard deviation for each channel. 32 channels of the feature maps were grouped for group normalization in the quality encoder. Since we utilized a two-stage hierarchy, a total of four CQT blocks were used, and the number of heads in MHSA was equally set to 4 ($n_h = 4$) for all blocks. The embedding dimension was set to 128 ($d_m = 128$), thus the dimension of the final feature f_p was 3072 ($24d_m$), and the latent space dimension was 2048 ($12d_m$) in MLP prediction head with 0.1 dropout rate. ADAM optimizer was used with a batch size 8, and a learning rate scheduler with cosine decay was applied with an initial learning rate of 0.0001. A mean absolute error was employed as a training loss function targeting MOS values. Because the fixed-size image has to be fed into our model, for testing, each image was subdivided 256 × 256 size overlapping patches, and the predicted scores were averaged. Our implementation was based on the PyTorch framework, and it took about six hours to train the model to achieve the desired performance (about 30 epochs) with a single NVIDIA RTX3090 GPU.

C. PERFORMANCE MEASURES

Since each dataset has a different subjective score range and distribution, the following well-known statistical measurements were used to benchmark the performance of previous IQA methods: Pearson linear correlation coefficient (PLCC) [61], Spearman rank correlation coefficient (SRCC) [62], and Kendall rank correlation coefficient (KRCC) [63]. In the experiment, each correlation coefficient is an average of values obtained over 30 training and testing iterations. The prediction of the proposed model is compared with several previous IQA methods. Here, all the Transformer-based models trained onto KADID-10k dataset. The results are tabulated in Table 1. As shown in the result, the proposed CQT delivers competitive performance in comparison to both conventional closed-form metrics and learning-based methods. In particular, the CQT outperforms the other learning-based methods across four databases. Overall, outstanding performances over each database are achieved with the Transformer-based models. The result demonstrates that the improved understandability of long-range dependency over spatial domain delivered by Transformer layers is effective to determine the level of visual quality. Whereas, the better performances over each database are achieved by the conventional metrics, for example, VIF [60] and FSIM$_c$ [7] for LIVE, GMSD [33] for CSIQ, and VSI [8] for TID2013 and TID2008 datasets. It is noteworthy that such well-known datasets have been re-used over several years throughout the design processes, and these FR-IQA methods might be intentionally over-adapted to certain subjective opinions and degradation patterns [14]. Otherwise, the proposed method shows consistent and outstanding performance for all datasets proving that the CQT is a good generalized IQA model. Figs. 4 (a)-(b) depict the correlation results for the image quality predicted by the CQT and the ground-truth MOS as scatter plots. As shown in these results, it is obvious that the predicted image quality scores are closely regressed onto the ground-truth, i.e., the proposed model is highly correlated with the subjective opinions over the different databases.

In order to demonstrate the proposed model well understands the representative quality factors in latent space to infer the degree of image quality, the pooled feature vector $f_p \in \mathbb{R}^{24d_m}$ prior to the MOS regression layer is visualized by being embedded into a lower dimension. Fig. 5 shows a generated two-dimensional manifold of CSIQ datasets obtained using t-SNE [64], where each point represents an image, and the points are labeled in accordance with the distorted type. Note that although none of the information associated with distortion type has been provided when training the model, the data points in the graphed manifold are clearly separated according to their distortion type. This indicates that the CQT is effectively trained to predict image quality with reasoning degradation and extracting meaningful features.

TABLE 1. PLCC, SRCC and KRCC comparison on the four databases including LIVE [28], CSIQ [29], TID2013 [30], and TID2008 [31]. The top three values are highlighted as bold face.

Method		LIVE [28]			CSIQ [29]			TID2013 [30]			TID2008 [31]		
		PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SROCC	KRCC
Conventional Metrics	PSNR	0.865	0.873	0.680	0.819	0.810	0.601	0.677	0.687	0.496	0.489	0.525	0.393
	SSIM [2]	0.937	0.948	0.796	0.852	0.865	0.680	0.777	0.727	0.545	0.600	0.624	0.452
	MS-SSIM [59]	0.940	0.951	0.805	0.889	0.906	0.730	0.830	0.786	0.605	0.789	0.853	0.655
	VSI [8]	0.948	0.952	0.806	0.928	0.942	0.786	**0.900**	**0.897**	**0.718**	0.864	**0.895**	**0.707**
	VIF [60]	**0.960**	**0.964**	0.828	0.913	0.911	0.743	0.771	0.677	0.518	0.776	0.749	0.586
	FSIM [7]	**0.961**	**0.965**	**0.836**	0.919	0.931	0.769	0.877	0.851	0.667	0.862	0.876	0.688
	NLPD [32]	0.932	0.937	0.778	0.923	0.932	0.769	0.839	0.800	0.625	-	-	-
	GMSD [33]	0.957	0.960	0.827	0.945	**0.950**	**0.804**	0.855	0.804	0.634	0.830	0.840	0.651
Learning-based Models	DeepIQA [34]	0.940	0.947	0.791	0.901	0.909	0.732	0.834	0.831	0.631	-	-	-
	PieAPP [36]	0.908	0.919	0.750	0.877	0.892	0.715	0.859	0.876	0.683	0.477	0.509	0.367
	LPIPS [12]	0.934	0.932	0.765	0.896	0.876	0.689	0.749	0.670	0.497	0.711	0.715	0.522
	DISTS [14]	0.954	0.954	0.811	0.928	0.929	0.767	0.855	0.830	0.639	0.830	0.808	0.619
	IQT [19]	0.938	0.937	0.788	0.898	0.897	0.730	0.868	0.848	0.657	**0.882**	0.875	0.690
	MSFPT [23]	0.950	0.952	0.817	**0.945**	0.938	0.780	0.878	0.866	0.675	**0.904**	**0.908**	**0.732**
	VTAMIQ [22]	**0.967**	**0.964**	**0.836**	**0.970**	**0.970**	**0.849**	**0.894**	**0.886**	**0.705**	0.861	0.863	0.684
	CQT (ours)	0.951	0.958	**0.833**	**0.962**	**0.964**	**0.830**	**0.896**	**0.888**	**0.706**	**0.887**	**0.883**	**0.693**

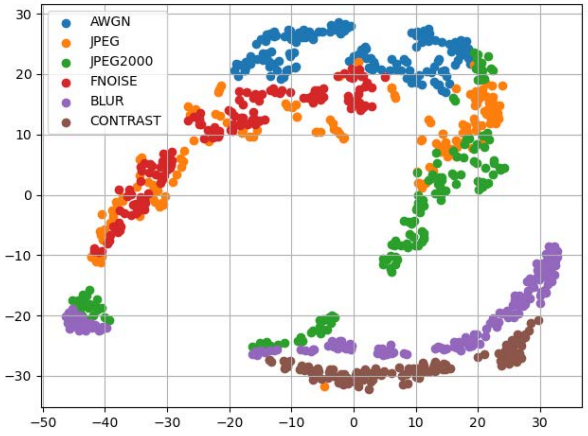


FIGURE 5. Visualization of two-dimensional manifold obtained by t-SNE. The manifold is projected from the extracted features (i.e., f_p in Fig. 1) of CSIQ dataset through the CQT. Each point indicates an image, and the points labeled according to the distorted type.

Fig. 6 depicts the examples of attention maps from the proposed CQT model. The left is the reference and distorted images pair in order. The images in the upper right and lower right are the attention maps obtained through the first and second stages, respectively. At each row, the four obtained attention maps are visualized since there are four multi-heads in each CQT block. Here, each attention map is averaged over all attention weights of each head. As shown in the figure, although there are some overlapped regions that the attention maps focus on, each attention map looks to play a different role from others for inference of quality. For example, in (a)-(c) which are globally distorted images by impulse noise, change of color saturation, and Gaussian blur, respectively. Some of the attention maps have higher weights on the monotonous regions where the artifacts are perceived more easily by humans than those of the high-frequency regions. Whereas, some attention maps concentrate on the textured regions which are not dealt with by the others, and whose distortions may also affect the perceptually degraded holistic quality. Moreover, some of the attention maps show that the proposed model also considers understanding the higher-level semantics to predict the degree of quality. (d)-(f) show locally distorted images by JPEG2000 compression, local block-wise distortion, and masked noise. In such cases, the attention map groups are more clearly separated where some of them focus on the local distorted regions and the others, and vice versa.

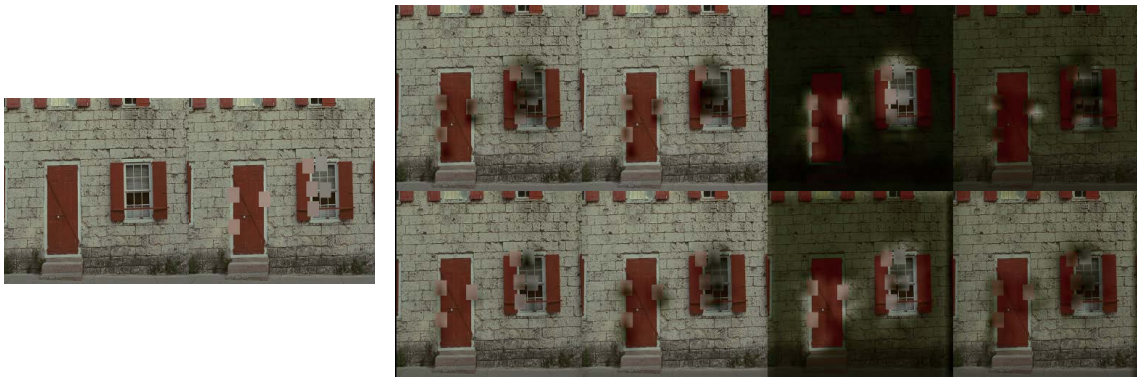
Because the proposed model consists of a multi-stage hierarchy, we observed that the attention maps at each stage differently resolve an image at scale. That is, multi-heads in the second stage more globally attend the image regions. In a way, it is obvious since the receptive field for LQPUs is wider than those of the first layer, and the spatial feature size is smaller which means MHSA deals the fewer tokens and each token interacts with the others in the longer range. Such visualization proves that the proposed CQT works correctly and whose each component plays its own role to predict image quality.

D. ABLATIONS

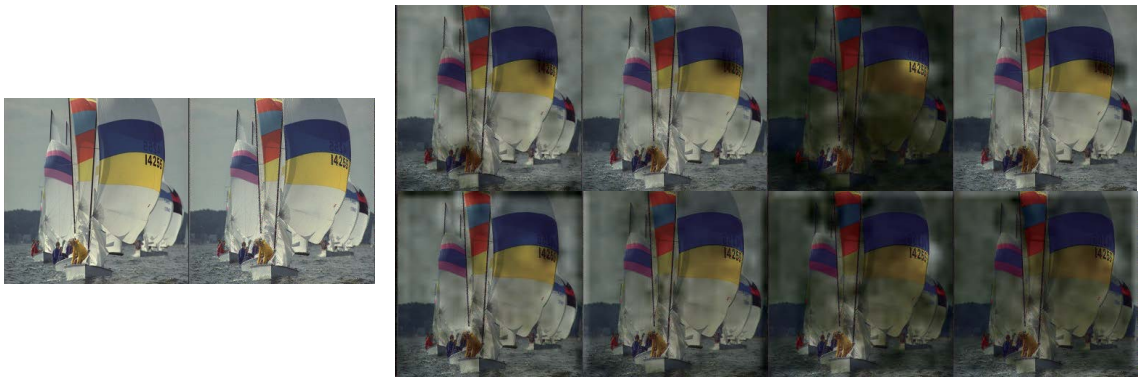
1) EFFECT OF ENCODERS

Two types of projectors are utilized in the proposed model including the quality and semantic encoders as aforementioned in the previous Section III-A. In order to achieve the best predictive performance, we conducted ablation experiments to investigate the effect of the encoder. Four kinds of scenarios are considered and whose performances are tabulated in Table 2. The semantic encoder (i.e., pre-trained Inception-ResNet-V2, encoder type B) for downsampling an image showed better performance rather than employing the 8×8 patch tokenization approach as used in ViT (encoder type A) [16]. When the frozen network was substituted by the learnable encoder (i.e., the quality encoder, encoder type C), the predictive power was slightly improved. As shown in the results, employing both the quality and semantic encoders had the best performance, thus encoder type D was chosen as the final model.

FIGURE 6. Visualized attention map examples (please zoom in to see details). Left side images show the reference and distorted image pair. Distortion types are (a) impulse noise, (b) change of color saturation, (c) Gaussian blur, (d) JPEG2000 compression, (e) local block-wise distortion, and (f) masked noise. The images on the right side are attention maps obtained from the stage 1 (upper) and 2 (lower), respectively.



(e)



(f)

FIGURE 6. ***(Continued.)*** **Visualized attention map examples (please zoom in to see details). Left side images show the reference and distorted image pair. Distortion types are (a) impulse noise, (b) change of color saturation, (c) Gaussian blur, (d) JPEG2000 compression, (e) local block-wise distortion, and (f) masked noise. The images on the right side are attention maps obtained from the stage 1 (upper) and 2 (lower), respectively.**

TABLE 2. **Performance comparison on four standard IQA databases depending on the type of image encoder.**

	Encoder type	LIVE			CSIQ			TID2013			TID2008		
		PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
A	8×8 patch tokenize	0.892	0.933	0.788	0.917	0.925	0.765	0.800	0.792	0.611	0.797	0.787	0.600
B	Semantic encoder only	0.894	0.941	0.798	0.952	0.951	0.809	0.891	0.885	0.698	0.881	0.880	0.688
C	Quality encoder only	0.925	0.950	0.814	0.953	0.963	0.827	0.889	0.885	0.696	0.882	0.881	0.690
D	Quality encoder + semantic encoder (proposed)	0.951	0.958	0.833	0.962	0.964	0.830	0.896	0.888	0.706	0.887	0.883	0.693

2) EFFECT OF MULTI-STAGES

We investigated whether the multi-scale feature extraction can lead the improved predictive power on image quality or not, and how many stages are optimal. Towards this, the ablation studies were performed with varying the number of stages. At first, a performance verification was started from a single-stage setting, and then an additional stage was stacked one by one. Here, we set the spatial resolution of the feature map to be reduced in half at each moving on to the next stage. The extracted feature embeddings from the third stage were also pooled by (6) and (7), and which were concatenated with f_1 and f_2 to be fed into the MLP prediction head. As tabulated in Table 3, when the two stages hierarchy was employed (case B), the model's performance was overall improved rather than the single-stage architecture (case A). However, the predictive power decreased when the three stages were utilized (case C), demonstrating that it is difficult to guarantee that more stages make the model have better performance. Fig. 7 visualizes attention maps derived by MHSA of stage 3 in case C. It can be analyzed that the spatial size of the feature map (i.e., $\frac{h}{16} \times \frac{w}{16}$ tokens) belonging to stage 3 is excessively small to extract additional information from local regions. Consequently, stage 3 for understanding an additional relationship between the uninformative local perceptions is unnecessary as shown in Fig. 7, and this might be the reason why case C showed the decreased performance.

3) EFFECT OF QUALITY POOLING

The proposed model applied average and p-percentile poolings ((6) and (7)) to feature embeddings f_1 and f_2 to abstract higher-level information representing global image quality. To investigate the performance variation in accordance with

TABLE 3. Performance comparison on four standard IQA databases depending on the number of stages considering multi-stage hierarchy.

	The number of stages	LIVE			CSIQ			TID2013			TID2008		
		PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
A	Single stage	0.923	0.932	0.783	0.954	0.952	0.814	0.893	0.889	0.694	0.874	0.869	0.686
B	2 stages (proposed)	0.951	0.958	0.833	0.962	0.964	0.830	0.896	0.888	0.706	0.887	0.883	0.693
C	3 stages	0.934	0.940	0.807	0.939	0.945	0.796	0.890	0.882	0.682	0.863	0.860	0.677

FIGURE 7. An example shows redundant information causes degradation of performance. The left side image is a reference and distorted image pair. Distortion type is spatially correlated noise. The images on the right side are the visualized attention maps of stage 3 when three stages were employed. In this case, the MOS of the distorted image is 0.324, but the predicted quality score is 0.760.

TABLE 4. Performance comparison on four standard IQA databases depending on the pooling methods to abstract global quality from feature embeddings.

	Pooling methods	LIVE			CSIQ			TID2013			TID2008		
		PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC	PLCC	SRCC	KRCC
A	Average pooling only	0.923	0.925	0.771	0.942	0.939	0.788	0.887	0.882	0.692	0.871	0.859	0.670
B	p-percentile pooling only	0.944	0.935	0.816	0.955	0.961	0.824	0.894	0.881	0.698	0.886	0.878	0.679
C	Average pooling + p-percentile pooling (proposed)	0.951	0.958	0.833	0.962	0.964	0.830	0.896	0.890	0.706	0.887	0.883	0.693
D	C+ variance pooling	0.945	0.939	0.819	0.957	0.959	0.823	0.887	0.869	0.680	0.873	0.865	0.672

the different pooling schemes, four methods were tested as tabulated in Table 4. We observed that the p-percentile pooling (method B) led the improved performance rather than the average pooling (method A) when only a single pooling method was applied. Our proposed method (method C) utilized both poolings together which make the model achieve better performance. Here, a doubt that arises here was whether more statistics can help increase predictive power or not. Thus, we tested method D where an additional pooling was added to method C which estimates variance over the feature embeddings along with the spatial domain as:

$$p_v(f) = \frac{1}{N_t}\sum_{i=1}^{N_t}\left(f_{i,j}^h - p_m(f)\right)^2, \tag{8}$$

where notations are the same as in (6) and (7). This aimed for reflecting how much the distortion is spatially dispersed [65], however, as presented in Table 4, the performance even deteriorated when an additional pooling method was applied. This indicates that simply adding an aggregation scheme is unimportant and the utilized poolings sufficiently capture global information to predict the degree of image quality.

V. CONCLUSION

A data-driven FR-IQA model CQT taking a hybrid architecture consisting of CNN and Transformer has been proposed in this study. Since capturing representations of local distortion and estimating global image quality based on their long-range dependency is essential to reliable IQA, the proposed model aimed for taking advantage of both CNN and Transformer regarding quality perception. Towards achieving the generalized predictive performance, we designed substructures of CQT to reflect HVS and explored their optimal combination by observing the interaction between local perceptions within each stage. The CQT demonstrated the improved and generalized performance over three standard datasets in comparison with the several previous IQA schemes. Moreover, we showed that the local feature extraction and the global quality abstraction processes work appropriately with complement each other by visualizing their interaction and by analyzing the performance varies according to model ablations. Nevertheless, the proposed method has limitations in that it requires pristine reference to predict the degree of visual quality. In addition, recent studies are focusing on the subjective hallucinations led by generative image restoration models, but we dealt with conventional artifacts. Hence, now we are attempting to construct a general-purpose IQA model without reference for more practical services since recent NR-IQA models still lack performance compared to FR-IQA level. If we design a powerful image restoration network synthesizing a pristine image from the distorted one, it would be possible for the similar approach proposed in this work to be migrated to NR-IQA scenario. Beyond image quality, only a few studies of an objective quality metric for the rendered 3D scene exist, thus we also intend to form a database for quantifying 3D quality and analyzing the human's visual experiences when viewing 3D objects.

REFERENCES

[1] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[2] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: From error visibility to structural similarity," *IEEE Trans. Image Process.*, vol. 13, no. 4, pp. 600–612, Apr. 2004.

[3] Z. Wang, G. Wu, H. R. Sheikh, E. P. Simoncelli, E.-H. Yang, and A. C. Bovik, "Quality-aware images," *IEEE Trans. Image Process.*, vol. 15, no. 6, pp. 1680–1689, Jun. 2006.

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

[5] Z. Wang, Bovik, and L. Lu, "Why is image quality assessment so difficult?" in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, May 2002, p. 3313.

[6] W.-J. Tsai and Y.-S. Liu, "Foveation-based image quality assessment," in *Proc. IEEE Vis. Commun. Image Process. Conf.*, Dec. 2014, pp. 25–28.

[7] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE Trans. Image Process.*, vol. 20, no. 8, pp. 2378–2386, Aug. 2011.

[8] L. Zhang, Y. Shen, and H. Li, "VSI: A visual saliency-induced index for perceptual image quality assessment," *IEEE Trans. Image Process.*, vol. 23, no. 10, pp. 4270–4281, Aug. 2014.

[9] P. Zhang, W. Zhou, L. Wu, and H. Li, "SOM: Semantic obviousness metric for image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 2394–2402.

[10] Y. Liang, J. Wang, Y. Gong, and N. Zheng, "Image quality assessment using similar scene as reference," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2016, pp. 3–18.

[11] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu, "DeepSim: Deep similarity for image quality assessment," *Neurocomputing*, vol. 257, pp. 104–114, Feb. 2017.

[12] R. Zhang, P. Isola, A. A. Efros, E. Shechtman, and O. Wang, "The unreasonable effectiveness of deep features as a perceptual metric," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 586–595.

[13] L.-H. Chen, C. G. Bampis, Z. Li, and A. C. Bovik, "Learning to distort images using generative adversarial networks," *IEEE Signal Process. Lett.*, vol. 27, pp. 2144–2148, 2020.

[14] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2020.

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

[16] A. Dosovitskiy, L. Beyer, K. Alexander, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, J. Uszkoreit, and N. Houlsby, "An image is worth 16×16 words: Transformers for image recognition at scale," in *Proc. Int. Conf. Learn. Represent. (ICLR)*, May 2021.

[17] J. Jain, A. Singh, N. Orlov, Z. Huang, J. Li, S. Walton, and H. Shi, "SeMask: Semantically masked transformers for semantic segmentation," 2021, *arXiv:2112.12782*.

[18] H. Zhang, F. Li, S. Liu, L. Zhang, H. Su, J. Zhu, L. M. Ni, and H.-Y. Shum, "DINO: DETR with improved DeNoising anchor boxes for end-to-end object detection," 2022, *arXiv:2203.03605*.

[19] M. Cheon, S.-J. Yoon, B. Kang, and J. Lee, "Perceptual image quality assessment with transformers," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2021, pp. 433–442.

[20] J. You and J. Korhonen, "Transformer for image quality assessment," 2020, *arXiv:2101.01097*.

[21] W. Jiang, L. Li, Y. Ma, Y. Zhai, Z. Yang, and R. Wang, "Image quality assessment with transformers and multi-metric fusion modules," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2022, pp. 1805–1809.

[22] A. Chubarau and J. Clark, "VTAMIQ: Transformers for attention modulated image quality assessment," 2021, *arXiv:2110.01655*.

[23] A. Keshari, S. Komal, and B. Subudhi, "Multi-scale features and parallel transformers based image quality assessment," 2022, *arXiv:2204.09779*.

[24] H. Oh, S. Ahn, J. Kim, and S. Lee, "Blind deep S3D image quality evaluation via local to global feature aggregation," *IEEE Trans. Image Process.*, vol. 26, no. 10, pp. 4923–4936, Oct. 2017.

[25] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, pp. 436–444, Feb. 2015.

[26] A. Khan, A. Sohail, U. Zahoora, and A. S. Qureshi, "A survey of the recent architectures of deep convolutional neural networks," *Artif. Intell. Rev.*, vol. 53, pp. 5455–5516, Apr. 2020.

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

[28] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Trans. Image Process.*, vol. 15, no. 11, pp. 3441–3452, Nov. 2006.

[29] E. C. Larson and D. M. Chandler, "Most apparent distortion: Full-reference image quality assessment and the role of strategy," *J. Electron. Imag.*, vol. 19, no. 1, 2010, Art. no. 011006.

[30] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti, and C.-C. J. Kuo, "Image database TID2013: Peculiarities, results and perspectives," *Signal Process., Image Commun.*, vol. 30, pp. 55–77, Jan. 2015.

[31] N. Ponomarenko, V. Lukin, A. Zelensky, K. Egiazarian, J. Astola, M. Carli, and F. Battisti, "TID2008—A database for evaluation of full-reference visual quality assessment metrics," *Adv. Modern Radioelectronics*, vol. 10, pp. 30–45, Jan. 2009.

[32] V. Laparra, J. Ballé, A. Berardino, and E. P. Simoncelli, "Perceptual image quality assessment using a normalized Laplacian pyramid," *Electron. Imag.*, vol. 28, no. 16, pp. 1–6, Feb. 2016.

[33] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Trans. Image Process.*, vol. 23, no. 2, pp. 684–695, Feb. 2013.

[34] S. Bosse, D. Maniry, K. R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2017.

[35] J. Kim and S. Lee, "Deep learning of human visual sensitivity in image quality assessment framework," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1676–1684.

[36] E. Prashnani, H. Cai, Y. Mostofi, and P. Sen, "PieAPP: Perceptual image-error assessment through pairwise preference," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 1808–1817.

[37] M. Zhu, G. Hou, X. Chen, J. Xie, H. Lu, and J. Che, "Saliency-guided transformer network combined with local embedding for no-reference image quality assessment," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshops (ICCVW)*, Oct. 2021, pp. 1953–1962.

[38] J. Ke, Q. Wang, Y. Wang, P. Milanfar, and F. Yang, "MUSIQ: Multi-scale image quality transformer," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 5148–5157.

[39] L. Yuan, Y. Chen, T. Wang, W. Yu, Y. Shi, Z. Jiang, F. E. Tay, J. Feng, and S. Yan, "Tokens-to-Token ViT: Training vision transformers from scratch on ImageNet," 2021, *arXiv:2101.11986*.

[40] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," 2021, *arXiv:2103.14030*.

[41] X. Dong, J. Bao, D. Chen, W. Zhang, N. Yu, L. Yuan, D. Chen, and B. Guo, "CSWin transformer: A general vision transformer backbone with cross-shaped Windows," 2021, *arXiv:2107.00652*.

[42] C.-F. Chen, Q. Fan, and R. Panda, "CrossViT: Cross-attention multi-scale vision transformer for image classification," 2021, *arXiv:2103.14899*.

[43] A. Hassani, S. Walton, N. Shah, A. Abuduweili, J. Li, and H. Shi, "Escaping the big data paradigm with compact transformers," 2021, *arXiv:2104.05704*.

[44] T. Xiao, M. Singh, E. Mintun, T. Darrell, P. Dollár, and R. Girshick, "Early convolutions help transformers see better," 2021, *arXiv:2106.14881*.

[45] H. Wu, B. Xiao, N. Codella, M. Liu, X. Dai, L. Yuan, and L. Zhang, "CvT: Introducing convolutions to vision transformers," 2021, *arXiv:2103.15808*.

[46] B. Heo, S. Yun, D. Han, S. Chun, J. Choe, and S. J. Oh, "Rethinking spatial dimensions of vision transformers," 2021, *arXiv:2103.16302*.

[47] B. Graham, A. El-Nouby, H. Touvron, P. Stock, A. Joulin, H. Jégou, and M. Douze, "LeViT: A vision transformer in ConvNet's clothing for faster inference," 2021, *arXiv:2104.01136*.

[48] J. Guo, K. Han, H. Wu, Y. Tang, X. Chen, Y. Wang, and C. Xu, "CMT: Convolutional neural networks meet vision transformers," 2021, *arXiv:2107.06263*.

[49] P. Esser, R. Rombach, and B. Ommer, "Taming transformers for high-resolution image synthesis," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 12873–12883.

[50] A. S. Shamsabadi, R. Sanchez-Matilla, and A. Cavallaro, "ColorFool: Semantic adversarial colorization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 1148–1157.
[51] X. Wang, K. Yu, S. Wu, J. Gu, Y. Liu, C. Dong, Y. Qiao, and C. C. Loy, "ESRGAN: Enhanced super-resolution generative adversarial networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 63–79.
[52] R. Mechrez, I. Talmi, and L. Zelnik-Manor, "The contextual loss for image transformation with non-aligned data," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 768–783.
[53] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Proc. Assoc. Advancement Artif. Intell. (AAAI)*, Feb. 2017, pp. 4278–4284.
[54] Z. Dai, H. Liu, Q. V. Le, and M. Tan, "CoAtNet: Marrying convolution and attention for all data sizes," 2021, *arXiv:2106.04803*.
[55] S. Mehta and M. Rastegari, "MobileViT: Light-weight, general-purpose, and mobile-friendly vision transformer," 2021, *arXiv:2110.02178*.
[56] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a 'completely blind' image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Nov. 2012.
[57] A. K. Moorthy and A. C. Bovik, "Visual importance pooling for image quality assessment," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 193–201, Apr. 2009.
[58] H. Lin, V. Hosu, and D. Saupe, "KADID-10k: A large-scale artificially distorted IQA database," in *Proc. 11th Int. Conf. Qual. Multimedia Exper. (QoMEX)*, Jun. 2019.
[59] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, vol. 2, Nov. 2004, pp. 1398–1402.
[60] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.
[61] K. Pearson, "VII. Mathematical contributions to the theory of evolution.—III. Regression, heredity, and panmixia," *Philos. Trans. R. Soc. London*, vol. 187, pp. 253–318, Jan. 1896.
[62] C. Spearman, "The proof and measurement of association between two things," *Amer. J. Psychol.*, vol. 15, no. 1, pp. 72–101, Jan. 1904.
[63] M. Kendall, "A new measure of rank correlation," *Biometrika*, vol. 30, nos. 1–2, pp. 81–89, Jun. 1938.
[64] L. van der Maaten and G. Hinton, "Visualizaing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
[65] M. Wu, L. Chen, and J. Tian, "A hybrid learning-based framework for blind image quality assessment," *Multidimensional Syst. Signal Process.*, vol. 29, no. 3, pp. 839–849, Feb. 2017.

HEESEOK OH received the B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2010, 2012, and 2017, respectively. He was a Senior Engineer of Samsung Electronics, Seoul. From 2017 to 2022, he was with the Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea. He is currently an Assistant Professor with the Department of Applied AI, Hansung University, Seoul. His research interests include 2D/3D image and video processing based on human visual systems, computer vision, extended reality, and deep generative model.

JINWOO KIM received the B.S. degree in electrical and electronic from Hongik University, Seoul, South Korea, in 2016. He is currently pursuing the M.S. and Ph.D. degrees with the Multi-Dimensional Insight Laboratory, Yonsei University. His current research interests include low-level computer vision, 3D reconstruction, perceptual image and video processing, and generative models for image, video, motion, and audio.

TAEWAN KIM received the B.S., M.S., and Ph.D. degrees in electrical and electronic engineering from Yonsei University, Seoul, South Korea, in 2008, 2010, and 2015, respectively. From 2015 to 2021, he was with the Vision AI Research Group, SK Telecom, Seoul. In 2022, he joined the Faculty of the Division of Future Convergence (Data Science Major), Dongduk Women's University, Seoul, where he is currently an Assistant Professor. His research interests include computer vision, and machine learning, including continual and online learning.

SANGHOON LEE (Senior Member, IEEE) received the B.S. degree from Yonsei University, South Korea, in 1989, the M.S. degree from the KAIST, South Korea, in 1991, and the Ph.D. degree from The University of Texas at Austin, Austin, TX, USA, in 2000. From 1991 to 1996, he was with Korea Telecom, South Korea. From 1999 to 2002, he was with Lucent Technologies, Murray Hill, NJ, USA. In 2003, he joined the Department of EE, Yonsei University, as a Faculty Member, where he is currently a Full Professor. His current research interests include image/video processing, computer vision, and graphics. He was a member of the IEEE IVMSP/MMSP TC, from 2014 to 2019 and from 2016 to 2021, respectively. He is a BoG Member of APSIPA. He was the General Chair of the 2013 IEEE IVMSP Workshop. Since 2011, he has been serving as the Chair of the IEEE P3333.1 Working Group. He was the Image, Video, and Multimedia TC Chair of APSIPA, from 2018 to 2019. He also served as an Editor for the *Journal of Communications and Networks*, from 2009 to 2015. He was an Associate Editor and a Guest Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, from 2010 to 2014, and 2013, respectively. He served as an Associate Editor and has been serving as a Senior Area Editor for the IEEE SIGNAL PROCESSING LETTERS, from 2014 to 2018, and since 2018. He has been serving as an Associate Editor for the IEEE TRANSACTIONS ON MULTIMEDIA, since 2022. He is also the Editor-in-Chief of APSIPA News Letters.