Data-Efficient Image Quality Assessment with Attention-Panel Decoder

Guanyi Qin[1*], Runze Hu[2*], Yutao Liu[3], Xiawu Zheng[4,5], Haotian Liu[1], Xiu Li[1†], Yan Zhang[5†],

[1] Tsinghua Shenzhen International Graduate School, Tsinghua University, Shenzhen 518055, China
[2] School of Information and Electronics, Beijing Institute of Technology, Beijing 100086, China
[3] School of Computer Science and Technology, Ocean University of China, Qingdao 266100, China
[4] Peng Cheng Laboratory, Shenzhen 518066, China
[5] Media Analytics and Computing Lab, Department of Artificial Intelligence, School of Informatics, Xiamen University, Xiamen 361005, China
{qgy21, liu-ht21}@mails.tsinghua.edu.cn, {hrzlpk2015, bzhy986}@gmail.com, liuyutao@ouc.edu.cn, zhengxw01@pcl.ac.cn, li.xiu@sz.tsinghua.edu.cn

Abstract

Blind Image Quality Assessment (BIQA) is a fundamental task in computer vision, which however remains unresolved due to the complex distortion conditions and diversified image contents. To confront this challenge, we in this paper propose a novel BIQA pipeline based on the Transformer architecture, which achieves an efficient quality-aware feature representation with much fewer data. More specifically, we consider the traditional fine-tuning in BIQA as an interpretation of the pre-trained model. In this way, we further introduce a Transformer decoder to refine the perceptual information of the CLS token from different perspectives. This enables our model to establish the quality-aware feature manifold efficiently while attaining a strong generalization capability. Meanwhile, inspired by the subjective evaluation behaviors of human, we introduce a novel attention panel mechanism, which improves the model performance and reduces the prediction uncertainty simultaneously. The proposed BIQA method maintains a lightweight design with only one layer of the decoder, yet extensive experiments on eight standard BIQA datasets (both synthetic and authentic) demonstrate its superior performance to the state-of-the-art BIQA methods, i.e., achieving the SRCC values of 0.875 (vs. 0.859 in LIVEC) and 0.980 (vs. 0.969 in LIVE). Checkpoints, logs and code will be available at https://github.com/narthchin/DEIQT.

Introduction

The goal of Image Quality Assessment (IQA) approaches is to automatically evaluate the quality of images in accordance with human subjective judgement. With the increasing growth of computer vision applications, the efficient and reliable IQA model has increased in importance. It is essential to monitor and improve the visual quality of contents and can be also adopted as testing criteria or optimization goals for benchmarking image processing algorithms. Based on the availability of the pristine reference image, IQA can be typically divided into full-reference IQA (FR-IQA) (Wang et al. 2004), reduced-reference IQA (RR-IQA) (Soundararajan and Bovik 2011), and no-reference or blind IQA (BIQA) (Moorthy and Bovik 2011). The applications of FR and RR IQA methods tend to be limited, since reference images are generally unavailable in real-world situations. Correspondingly, the BIQA methods do not require such reference images and thus become more promising yet more challenging.

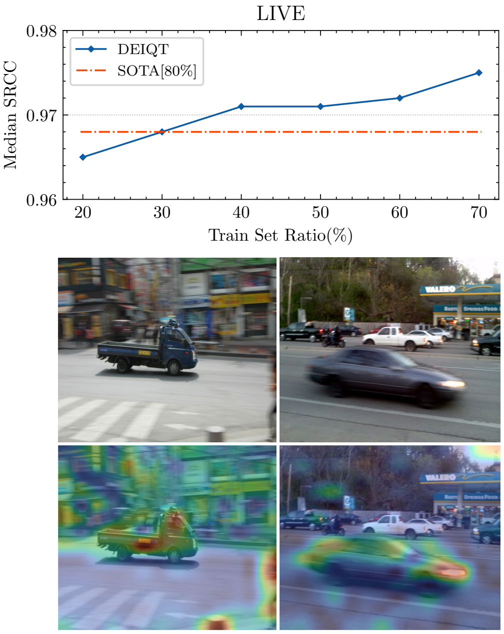

Figure 1: Image on top: the performance of the proposed DEIQT varying the amount of training data on the LIVE dataset. SOTA results are obtained from the TReS (Golestaneh, Dadsetan, and Kitani 2022) using 80% data. Our method can achieve the SOTA performance with only 30% data. Images in the medium: the sample images. Images at bottom: Quality attention map from DEIQT. Our model can accurately capture the quality degradation areas of an image. Meanwhile, it ignores the perceptual information that is related to the image recognition yet less important for the quality assessment, i.e., the white cars in the second image.

Current state-of-the-art (SOTA) BIQA methods employ

[*] Equal contribution, † Corresponding author.

either convolutional neural networks (CNN) or Vision Transformer (ViT) based architectures (Dosovitskiy et al. 2021), which perform an end-to-end optimization of feature engineering and quality regression, simultaneously. The training strategy of BIQA methods generally follows a straightforward pre-training and fine-tuning pipeline. In the pre-training stage, models are trained on a large-scale classification dataset, i.e., ImageNet (Deng et al. 2009). Then, models are fine-tuned on a small-scale BIQA dataset. Nevertheless, the requirements of feature representations for these two stages are not consistent. The pre-training stage concentrates on the global semantic features that are highly related to the image content, whereas the fine-tuning stage needs to consider both global semantics and local details of an image (Raghu et al. 2021). Consequently, the process of fine-tuning still necessitates a substantial amount of data so as to successfully adapt the model awareness from the image content understanding to the image quality. However, due to the labor-intensive characteristics of image annotation, BIQA has high expectations for fitness on low data volumes. Thus, an efficient data-learning strategy, which is capable of constructing an accurate quality-aware feature manifold using a small quantity of data, is desired and has become a beneficial endeavor for computer vision tasks and industrial applications.

To this end, we propose a novel BIQA method that can efficiently characterize the image quality using much fewer data than existing BIQA methods. The proposed BIQA method is based on the Transformer encoder-decoder architecture, herein namely data-efficient image quality transformer (DEIQT). Specifically, we consider that learned features at the pre-training stage are highly related yet more abstract for the BIQA task. In other words, the fine-tuning from the classification task to the BIQA task can be regarded as an interpretation of feature abstractness. Based on this, the classification (CLS) token in the Transformer encoder is an abstract characterization of quality-aware features (Touvron et al. 2021b). Thus, it may not effectively develop an optimal feature representation for the image quality during the process of fine-tuning. To address this issue, we introduce the Transformer decoder to further decode the CLS token, thereby effectively adapting the token for the BIQA task.

In particular, we make use of the self-attention and cross-attention operations in the decoder to realize an optimal feature representation for the image quality. The self-attention decodes the aggregated features in the CLS token. It can diminish the significance of those features that are less relevant to the image quality. The resulting outputs of self-attention are handled as the query to the decoder, which is therefore more sensitive to quality-aware image features. The cross-attention performs the interactions between the query and the extracted features from the encoder. It refines the decoder embeddings such that making the extracted features highly related to the image quality. The Transformer decoder brings in an efficient learning property for DEIQT. This not only allows the DEIQT to accurately characterize the image quality using significantly fewer data (Fig. 1), but also improves the model training efficiency (Fig. 5). Notably, one layer decoder is adequate to deliver a satisfactory performance for DEIQT (Table 9), which ensures a lightweight design of our model.

Furthermore, due to the considerable variation in the image contents and distortion conditions, existing BIQA methods generally suffer from a high prediction uncertainty. This hinders the model stability, leading to an inaccurate prediction. To address this issue, we design a novel attention-panel mechanism in the decoder. This mechanism is inspired by the subjective evaluation system, wherein an image is scored by a number of participants and the mean of scores (MOS) is considered the label of this image. During the subjective quality evaluation, opinions of humans on an image differ from person to person. The attention-panel mechanism mimics such human behaviors by randomly initializing and further learning the opinion of each "human" on the image quality. Specifically, it makes use of the cross-attention of decoder to evaluate the image quality from different perspectives and concludes the quality evaluation based on the results from all of these perspectives. The attention-panel mechanism can improve the model stability while introducing almost zero parameters (Table 8).

In summary, contributions of this paper are the following:

- We make the first attempt to develop a BIQA solution based on the complete Transformer encoder-decoder architecture. We employ the CLS token as inputs to the decoder, to enable the proposed DEIQT to extract comprehensive quality-aware features from an image while attaining a high learning efficiency. To the best of our knowledge, we are the first to leverage the Transformer decoder for the IQA task.
- Inspired by the human subjective evaluation, we introduce a novel attention-panel mechanism to further improve the performance of DEIQT while reducing the prediction uncertainty. Notably, the attention-panel mechanism introduces almost no parameters to the model.
- We verify DEIQT on 8 benchmark IQA datasets involving a wide range of image contents, distortion types and dataset size. DEIQT outperforms other competitors across all these datasets.

Related Work

CNN-based BIQA. Benefiting from the powerful feature expression ability, CNN-based BIQA methods have gained a great deal of popularity recently (Ma et al. 2017; Zhang et al. 2018; Su et al. 2020; Bosse et al. 2017). One of the mainstreams of the CNN-based method (Kim and Lee 2016) is to integrate the feature learning and regression modeling into a general CNN framework so that developing an accurate and efficient quality representation. Modern CNN-based models (Zhang et al. 2018) also put great efforts into other perspectives of the BIQA challenges, i.e., the limited size of IQA dataset and complex distortion conditions.

In summary, CNN-based methods demonstrate great potential for BIQA tasks, but further efforts are required. Specifically, CNN-based methods usually adopt the image patches (Zhu et al. 2020; Su et al. 2020) as inputs or extract learned features from different layers of CNNs to form a multi-perceptual-scale representation, i.e., the shallow layer

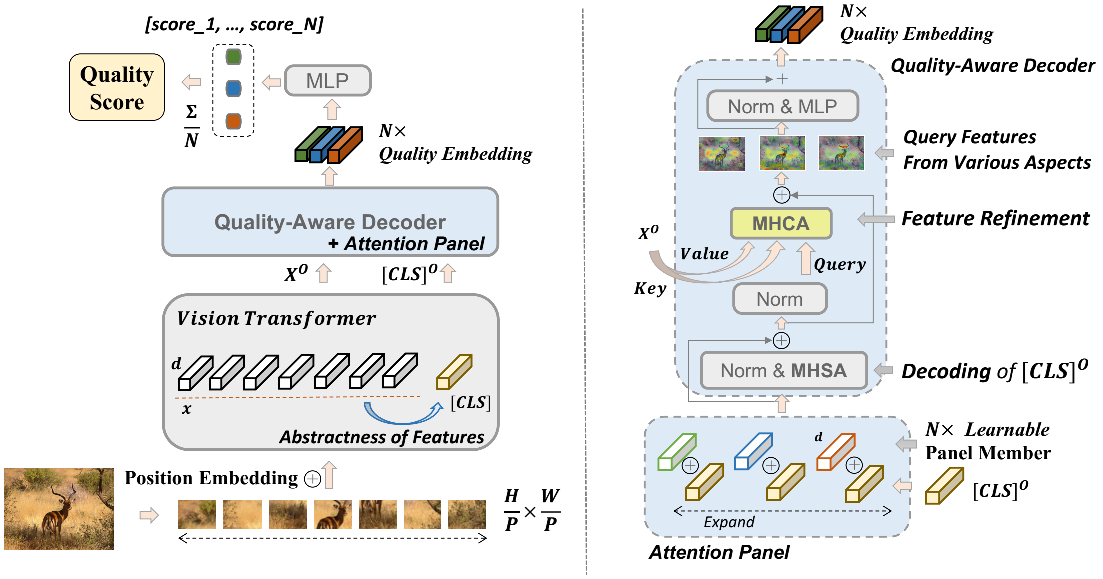

Figure 2: Model overview of the proposed DEIQT

for local details and the deeper layer for high-level semantics (Hu et al. 2021a,b). The effectiveness of these strategies has been proved, but it introduces non-negligible a computational burden in training and inference(Zhang et al. 2021, 2022). Furthermore, due to the inherent locality bias of CNNs, the CNN-based BIQA methods are often constrained by the ineffective characterization of non-local features, notwithstanding the fact that BIQA task depends on both local and non-local image information.

Transformers in BIQA. Transformers (Vaswani et al. 2017) that were first designed for the natural language processing have raised considerable research interests in the computer vision area. The Vision Transformer (ViT) (Dosovitskiy et al. 2021) is one of the most representative works. It performs the classification task using a pure Transformer encoder architecture, and, with modern training strategies, ViT achieves a competing performance against the CNN-based methods (Touvron et al. 2021a). Transformer also demonstrates great potential in dealing with the BIQA task thanks to its strong capability in modelling the non-local dependencies among perceptual features of the image. Currently, there are mainly two ways for using Transformer in BIQA: hybrid Transformer (Golestaneh, Dadsetan, and Kitani 2022; You and Korhonen 2021) and pure ViT-based Transformer (Ke et al. 2021). The former utilizes CNNs to extract the perceptual features as inputs to the Transformer encoder, whereas the latter directly sends image patches as inputs to the Transformer encoder.

The Transformer-based BIQA methods have achieved great performance. However, the Transformer in BIQA can be further exploited. Current Transformer-based BIQA methods only involve the Transformer encoder, yet their ability to accurately characterize the image quality is still restricted. The main reason can be attributed to that the extracted features from the encoder are rather abstract in terms of the image quality, making it difficult to model the relations between these features and the quality score. Thus, additional efforts are needed to derive an optimal feature representation for the image quality.

Data-Efficient Image Quality Transformer

Overall Architecture

To further improve the learning efficiency and capacity of BIQA, we make the first attempt to develop a Transformer encoder-decoder BIQA framework, namely data-efficient image quality transformer (DEIQT). The overall architecture of the proposed DEIQT is illustrated in Fig. 2. Given an input image, we first obtain the CLS token through the outputs of the Transformer encoder, which acts as the multi-perceptual-level image representation. With the self-attention operation, the CLS token can capture local and non-local dependencies from patch embeddings, thereby preserving comprehensive information for the image quality. The CLS token is then integrated with the attention-panel embeddings via the element-wise addition. A multi-head self-attention block is applied to transform them into queries of the decoder. Each attention-panel embedding absorbs the information from the CLS token, where the cross-attention mechanism in the decoder allows each to learn quality-aware features of an image from a unique perspective. Following this, the transformer decoder outputs the quality embeddings consisting of quality-aware features of an image. Finally, the quality embeddings are sent to a multi-layer perceptron (MLP) head to make several predictions for the image quality. We can obtain one prediction from each embedding of the quality embeddings. The average of these predictions is treated as the final quality score of the image.

Perceptual Feature Aggregation in Transformer Encoder

The self-attention of Transformer encoder aggregates local and non-local information from a sequence of input patches with a minimum inductive bias, which allows it to comprehensively characterize perceptual features of an image. We herein take the advantage of the self-attention to obtain an efficient perceptual representation for the image. Given an input image $I \in \mathbb{R}^{C \times H \times W}$, we first reshape it into N patches as in $t_n \in \mathbb{R}^{p^2 \times C}$, where H and W are the height and width of the image, respectively. C is the number of channels and p indicates the patch size. The total number of patches is calculated as in $N = \frac{HW}{p^2}$. Each patch is then transformed into a D-dimension embedding through a linear projection layer. A learnable embedding of CLS token $\boldsymbol{T}_{\text{CLS}} \in \mathbb{R}^{1 \times D}$ is prepended to the N patch embeddings yielding to a total number of $N+1$ embeddings. An additional position embedding is also introduced into these $N{+}1$ embeddings for preserving the positional information.

Let $\boldsymbol{T} = \{\boldsymbol{T}_{\text{CLS}}, \boldsymbol{T}_1, \ldots, \boldsymbol{T}_N\} \in \mathbb{R}^{N+1 \times D}$ be the embedding sequence. $\boldsymbol{T}$ is then fed to the multi-head self-attention (MHSA) block to perform the self-attention operation. The MHSA block contains h heads each with the dimension $d = \frac{D}{h}$. $\boldsymbol{T}$ is transformed into three groups of matrices as in the query $\boldsymbol{Q}$, key $\boldsymbol{K}$ and value $\boldsymbol{V}$ using three different linear projection layers, where $\boldsymbol{Q} = \{\boldsymbol{Q}_1, ..., \boldsymbol{Q}_h\} \in \mathbb{R}^{(N+1) \times D}$, $\boldsymbol{K} = \{\boldsymbol{K}_1, ..., \boldsymbol{K}_h\} \in \mathbb{R}^{(N+1) \times D}$, and $\boldsymbol{V} = \{\boldsymbol{V}_1, ..., \boldsymbol{V}_h\} \in \mathbb{R}^{(N+1) \times D}$ for $\boldsymbol{Q}_h, \boldsymbol{K}_h, \boldsymbol{V}_h \in \mathbb{R}^{(N+1) \times d}$. The output of Transformer encoder $\boldsymbol{Z}_O$ is formulated as :

$$\begin{aligned} \text{MHSA}\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) =& \text{Cat}(\textit{Attention}(\boldsymbol{Q}_1, \boldsymbol{K}_1, \boldsymbol{V}_1), \ldots, \\ & \textit{Attention}(\boldsymbol{Q}_h, \boldsymbol{K}_h, \boldsymbol{V}_h))\boldsymbol{\mathcal{W}}_L \\ \boldsymbol{Z}_M =& \text{MHA}\left(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}\right) + \boldsymbol{T} \\ \boldsymbol{Z}_O =& \text{MLP}\left(\text{Norm}(\boldsymbol{Z}_M)\right) + \boldsymbol{Z}_M, \end{aligned} \tag{1}$$

where $\boldsymbol{\mathcal{W}}_L$ refers to the weights of the linear projection layer, $\textit{Attention}(\boldsymbol{Q}_h, \boldsymbol{K}_h, \boldsymbol{V}_h) = \textit{softmax}\left(\frac{\boldsymbol{Q}_h \boldsymbol{K}_h{}^T}{\sqrt{d}}\right)\boldsymbol{V}_h$ and $\text{Norm}(\cdot)$ indicates the layer normalization. $\boldsymbol{Z}_O$ is denoted as in $\boldsymbol{Z}_O = \{\boldsymbol{Z}_O\left[0\right], ..., \boldsymbol{Z}_O\left[N\right]\} \in \mathbb{R}^{(N+1) \times d}$.

Quality-Aware Decoder

For the ViT-based BIQA methods, the learned CLS token $\boldsymbol{Z}_O\left[0\right]$ is typically considered to contain aggregated perceptual information for the image quality. It will be sent to an MLP head to perform the regression task of quality prediction. However, as explained earlier, $\boldsymbol{Z}_O\left[0\right]$ mainly relates to the abstractness of quality-aware features. It is difficult to directly utilize $\boldsymbol{Z}_O\left[0\right]$ to attain an optimal representation for the image quality. To this end, we introduce a quality-aware decoder to further interpret the CLS token, such that making the extracted features more significant to the image quality.

Let $\hat{\boldsymbol{T}}_{\text{CLS}} \in \mathbb{R}^{1 \times D}$ be the CLS token obtained from the output of encoder. $\hat{\boldsymbol{T}}_{\text{CLS}}$ is first sent to a MHSA block to model the dependencies between each element with the remaining elements of the CLS token. The output of MHSA is followed by the residual connection to generate queries of the transformer decoder, written by

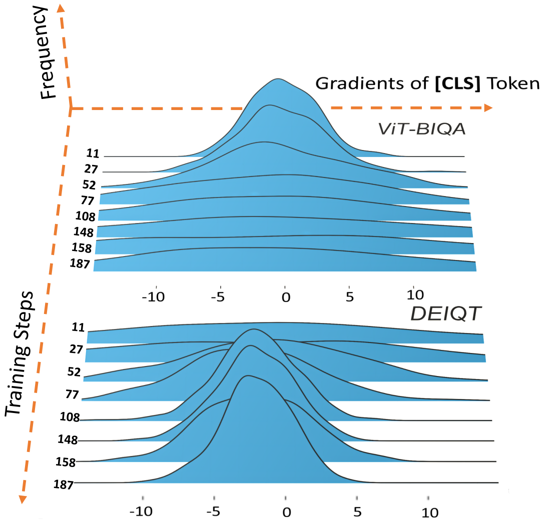

Figure 3: Probability distributions of CLS token Gradients varying the training steps. ViT-BIQA and DEIQT are models without and with the proposed decoder, respectively. By introducing the decoder, variations in the gradients decrease considerably faster than those without the decoder, indicating that the decoder can greatly improve training efficiency.

$$\boldsymbol{Q}_d = \text{MHSA}\left(\text{Norm}\left(\hat{\boldsymbol{T}}_{\text{CLS}}\right)\right) + \hat{\boldsymbol{T}}_{\text{CLS}}. \tag{2}$$

The role of the MHSA block is to decode the CLS token such that making the produced query more sensitive to the quality-aware features. Following this, we utilize $\hat{\boldsymbol{Z}}_O = \{\boldsymbol{Z}_O\left[1\right], ..., \boldsymbol{Z}_O\left[N\right]\} \in \mathbb{R}^{N \times d}$ as Key and Value of the decoder, denoted by $\boldsymbol{K}_d = \boldsymbol{V}_d = \hat{\boldsymbol{Z}}_O$, where $\hat{\boldsymbol{Z}}_O \cap \boldsymbol{Z}_O = \hat{\boldsymbol{T}}_{\text{CLS}}$. Then, $\boldsymbol{Q}_d, \boldsymbol{K}_d$ and $\boldsymbol{V}_d$ are sent to a multi-head cross-attention (MHCA) block to perform the cross-attention. During this process, we utilize $\boldsymbol{Q}_d$ to reinteract with the features of the image patches preserved in the encoder outputs, and thus ensuring the attentional features more significant to the image quality. The output of the cross-attention is written by

$$\boldsymbol{S} = \text{MLP}\left(\text{MHCA}(\text{Norm}(\boldsymbol{Q}_d), \boldsymbol{K}_d, \boldsymbol{V}_d) + \boldsymbol{Q}_d\right), \tag{3}$$

where $\boldsymbol{S}$ indicates the refined quality-aware features from the encoder outputs which is more comprehensive and accurate in defining the image quality. Finally, $\boldsymbol{S}$ is fed to an MLP head to derive the final quality score, wherein we minimize the smooth l_1 loss to train our network. The quality-aware decoder can significantly improve the learning capacity of the transformer-based BIQA model, and thus enhancing the model performance in terms of prediction accuracy, generalization capability and stability.

In Fig. 3, we demonstrate the effectiveness of the quality-aware decoder by investigating the gradients of the CLS token for models with and without the decoder. As observed, without the decoder, the gradients of the CLS token vary significantly throughout the training. This will substantially decrease the training efficiency, and even cause the model to fail to converge. Correspondingly, the designed decoder is capable of reducing such a large variation, thereby ensuring a high training efficiency. It is also worth mentioning that the designed Decoder is combined with the standard ViT encoder in a non-intrusive manner, which not only makes our model compatible with any variants of the ViT encoder but also enables us to directly utilize the weights of other pretrained encoders to increase the training efficiency of our model. More importantly, we empirically show that the designed Decoder with a depth of 1 can effectively achieve a satisfactory performance (Table 9). This significantly restricts the model size, making it more suitable for practical applications.

Attention-Panel Mechanism

Images captured in the real-world generally involve various contents and complex distortion conditions, resulting in the BIQA models exhibiting a high prediction uncertainty. To mitigate this, we propose an attention-panel mechanism in the Transformer decoder. This mechanism is inspired by the human subjective evaluation, wherein an image is scored by a number of participants and the mean of scores (MOS) is considered the label of the image. During this evaluation process, the personal subjective opinion on an image differs from person to person. The proposed attention-panel mechanism imitates such a situation, in which each panel member represents a participant of the subjection evaluation and judges the image quality from a different perspective. This way, the model can achieve a comprehensive evaluation of the image quality, thus reducing the prediction uncertainty (Hu et al. 2019).

Let L be the number of panel members. Prior to sending the CLS token to the decoder, we create the attention-panel embeddings as in $\boldsymbol{J} = \{\boldsymbol{J}_1, \ldots, \boldsymbol{J}_L\} \in \mathbb{R}^{L \times D}$. $\boldsymbol{J}$ is initialized with random numbers. Then, we expand the CLS token L times to form the matrix $\hat{\boldsymbol{T}} = \{\hat{\boldsymbol{T}}_{\text{CLS}}, \ldots, \hat{\boldsymbol{T}}_{\text{CLS}}\} \in \mathbb{R}^{L \times D}$. The element-wise summation of $\boldsymbol{J}$ and $\hat{\boldsymbol{T}}$ is employed as the inputs to the quality-aware decoder. Therefore, the calculation of queries in Eq. 2 is reformulated as

$$\hat{\boldsymbol{Q}}_d = \text{MHSA}\left(\text{Norm}\left(\hat{\boldsymbol{T}}_{\text{CLS}} + \boldsymbol{J}\right)\right) + \left(\hat{\boldsymbol{T}}_{\text{CLS}} + \boldsymbol{J}\right). \quad (4)$$

The operation of cross-attention is performed in Eq. 3 by replacing $\boldsymbol{Q}_d$ with $\hat{\boldsymbol{Q}}_d$. We obtain the quality embeddings $\hat{\boldsymbol{S}} = \{\hat{\boldsymbol{S}}_1, ...\hat{\boldsymbol{S}}_L\} \in \mathbb{R}^{L \times D}$. Finally, $\hat{\boldsymbol{S}}$ is sent to the MLP to derive a vector of scores as in $\boldsymbol{O} = \{O_1, \ldots, O_L\}$, which contains L scores corresponding to L members. The mean of these L scores $\frac{\sum_{l=1}^{L}}{L}$ is treated as final quality score.

With the attention-panel, DEIQT is capable of characterizing the image quality from different perspectives, thus attaining a comprehensive evaluation. To verify that, we adopt the cosine similarity metric to measure the similarity between the characterized perceptual features from any two panel members. Given an image, we obtain the quality embeddings from three trained DEIQT models with 6, 12 and 18 panel members, respectively. The cosine similarity between every two quality embeddings is computed. The results are reported in Fig. 4. As observed, the similarity between these panel members is extremely low. Accordingly, the quality-aware features described by each panel member are rather different.

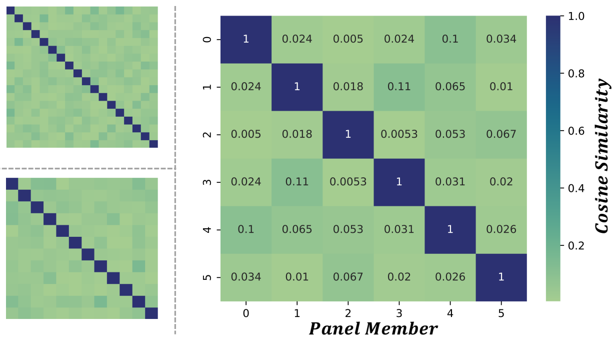

Figure 4: Cosine similarity between characterized perceptual features of panel members. The number of panel members in DEIQT is set to 6, 12, and 18, respectively. The extremely low similarity between two members suggests that each member judges the image quality from a very unique perspective.

Experiments

Benchmark Datasets and Evaluation Protocols

We evaluate the performance of the proposed DEIQT on 8 typical BIQA datasets, including 4 synthetic datasets of LIVE (Sheikh, Sabir, and Bovik 2006), CSIQ (Larson and Chandler 2010), TID2013 (Ponomarenko et al. 2015), KADID (Lin, Hosu, and Saupe 2019), and 4 authentic datasets of LIVEC (Ghadiyaram and Bovik 2015), KONIQ (Hosu et al. 2020), LIVEFB (Ying et al. 2020), SPAQ (Fang et al. 2020). Specifically, for the authentic datasets, LIVEC consists of 1162 images captured by different photographers with a wide variety of mobile devices. SPAQ contains 11000 images collected by 66 smartphones. KonIQ-10k is composed of 10073 images selected from public multimedia resources. LIVEFB is the largest-scale authentic dataset (by far) that includes 39810 images. For the synthetic datasets, they contain a small number of pristine images which are synthetically distorted by various distortion types, such as JPEG compression and Gaussian blurring. LIVE and CISQ contain 799 and 866 synthetically distorted images with 5 and 6 distortion types, respectively. TID2013 and KADID consist of 3000 and 10125 synthetically distorted images involving 24 and 25 distortion types, respectively.

In our experiments, two commonly used criteria, Spearman's rank order correlation coefficient (SRCC) and Pearson's linear correlation coefficient (PLCC), are adopted to quantify the performance of DEIQT in terms of prediction monotonicity and prediction accuracy, respectively. Both SRCC and PLCC range from 0 to 1. A superior performance should result in the absolute values of SRCC and PLCC close to 1.

	LIVE		CSIQ	
Method	PLCC	SRCC	PLCC	SRCC
DIIVINE	0.908	0.892	0.776	0.804
BRISQUE	0.944	0.929	0.748	0.812
ILNIQE	0.906	0.902	0.865	0.822
BIECON	0.961	0.958	0.823	0.815
MEON	0.955	0.951	0.864	0.852
WaDIQaM	0.955	0.960	0.844	0.852
DBCNN	**0.971**	0.968	**0.959**	**0.946**
TIQA	0.965	0.949	0.838	0.825
MetaIQA	0.959	0.960	0.908	0.899
P2P-BM	0.958	0.959	0.902	0.899
HyperIQA (*27M*)	0.966	0.962	0.942	0.923
TReS (*152M*)	0.968	**0.969**	0.942	0.922
MUSIQ (*27M*)	0.911	0.940	0.893	0.871
DEIQT (*24M*) (Ours)	**0.982**	**0.980**	**0.963**	**0.946**

Table 1: Performance comparison measured by medians of SRCC and PLCC on synthetic datasets of LIVE and CSIQ, where bold entries indicate the top two results.

	TID2013		KADID	
Method	PLCC	SRCC	PLCC	SRCC
DIIVINE	0.567	0.643	0.435	0.413
BRISQUE	0.571	0.626	0.567	0.528
ILNIQE	0.648	0.521	0.558	0.534
BIECON	0.762	0.717	0.648	0.623
MEON	0.824	0.808	0.691	0.604
WaDIQaM	0.855	0.835	0.752	0.739
DBCNN	0.865	0.816	0.856	0.851
TIQA	0.858	0.846	0.855	0.850
MetaIQA	0.868	0.856	0.775	0.762
P2P-BM	0.856	0.862	0.849	0.840
HyperIQA (*27M*)	0.858	0.840	0.845	0.852
TReS (*152M*)	**0.883**	**0.863**	0.858	0.859
MUSIQ (*27M*)	0.815	0.773	**0.872**	**0.875**
DEIQT (*24M*) (Ours)	**0.908**	**0.892**	**0.887**	**0.889**

Table 2: Performance comparison measured by medians of SRCC and PLCC on synthetic datasets of TID2013 and KADID, where bold entries indicate the top two results.

	LIVEC		KonIQ	
Method	PLCC	SRCC	PLCC	SRCC
DIIVINE	0.591	0.588	0.558	0.546
BRISQUE	0.629	0.629	0.685	0.681
ILNIQE	0.508	0.508	0.537	0.523
BIECON	0.613	0.613	0.654	0.651
MEON	0.710	0.697	0.628	0.611
WaDIQaM	0.671	0.682	0.807	0.804
DBCNN	0.869	0.851	0.884	0.875
TIQA	0.861	0.845	0.903	0.892
MetaIQA	0.802	0.835	0.856	0.887
P2P-BM	0.842	0.844	0.885	0.872
HyperIQA (*27M*)	**0.882**	**0.859**	0.917	0.906
TReS (*152M*)	0.877	0.846	**0.928**	0.915
MUSIQ (*27M*)	0.746	0.702	**0.928**	**0.916**
DEIQT (*24M*) (Ours)	**0.894**	**0.875**	**0.934**	**0.921**

Table 3: Performance comparison measured by medians of SRCC and PLCC on authentic datasets of LIVEC and KonIQ, where bold entries indicate the top two results.

	LIVEFB		SPAQ	
Method	PLCC	SRCC	PLCC	SRCC
DIIVINE	0.187	0.092	0.600	0.599
BRISQUE	0.341	0.303	0.817	0.809
ILNIQE	0.332	0.294	0.712	0.713
BIECON	0.428	0.407	-	-
MEON	0.394	0.365	-	-
WaDIQaM	0.467	0.455	-	-
DBCNN	0.551	0.545	0.915	0.911
TIQA	0.581	0.541	-	-
MetaIQA	0.507	0.540	-	-
P2P-BM	0.598	0.526	-	-
HyperIQA (*27M*)	0.602	0.544	0.915	0.911
TReS (*152M*)	0.625	0.554	-	-
MUSIQ (*27M*)	**0.661**	**0.566**	**0.921**	**0.918**
DEIQT (*24M*) (Ours)	**0.663**	**0.571**	**0.923**	**0.919**

Table 4: Performance comparison measured by medians of SRCC and PLCC on authentic datasets of LIVEFB and SPAQ, where bold entries indicate the top two results.

Implementation Details

For DEIQT, we followed the typical training strategy to randomly crop an input image into 10 image patches with a resolution of 224×224. Each image patch was then reshaped into a sequence of patches with the patch size $p = 16$, and the dimensions of input tokens $D = 384$. We created the Transformer encoder based on the ViT-S proposed in DeiT III (Touvron, Cord, and Jégou 2022). The depth of the encoder was set to 12, and the number of heads $h = 6$. For the Decoder, the depth was set to 1 and the number of panel members $L = 6$.

The Encoder of DEIQT was pre-trained on the ImageNet-1K for 400 epochs using the Layer-wise Adaptive Moments optimizer (You et al. 2020) for Batch training with the batch size 2048. DEIQT was trained for 9 Epochs. The learning rate was set to 2×10^{-4} with a decay factor of 10 every 3 epochs. The batch size was determined depending on the size of the dataset, i.e., 16 and 128 for the LIVEC and KonIQ, respectively. For each dataset, 80% images were used for training and the remaining 20% images were utilized for testing. We repeated this process 10 times to mitigate the performance bias and the medians of SRCC and PLCC were reported.

Overall Prediction Performance Comparison

Tables 1-4 report the comparison results between the proposed DEIQT and 13 state-of-the-art BIQA methods, which include both hand-crafted BIQA methods, such as DIIVINE (Saad, Bovik, and Charrier 2012), ILNIQE (Zhang, Zhang, and Bovik 2015) and BRISQUE (Mittal, Moorthy, and Bovik 2012), and deep-learning-based methods, i.e., MUSIQ (Ke et al. 2021) and MetaIQA (Zhu et al. 2020). As observed across these eight datasets, DEIQT outperforms all other methods. Since images on these 8 datasets span a wide

Training	LIVEFB		LIVEC	KonIQ	LIVE	CSIQ
Testing	KonIQ	LIVEC	KonIQ	LIVEC	CSIQ	LIVE
DBCNN	0.716	0.724	0.754	0.755	0.758	0.877
P2P-BM	0.755	0.738	0.740	0.770	0.712	-
HyperIQA	**0.758**	0.735	**0.772**	0.785	0.744	0.926
TReS	0.713	0.740	0.733	0.786	0.761	-
DEIQT	0.733	**0.781**	0.744	**0.794**	**0.781**	**0.932**

Table 5: SRCC on the cross datasets validation. The best performances are highlighted with boldface.

variety of image contents and distortion types, it is very challenging to consistently achieve the leading performance on all of them. Correspondingly, these observations confirm the effectiveness and superiority of DEIQT in characterizing the image quality.

Generalization Capability Validation

We further evaluate the generalization capability of DEIQT through the cross-datasets validation methodology, where a BIQA model is trained on one dataset, and then tested on the other datasets without any fine-tuning or parameter adaption. The experimental results in terms of the medians of SRCC on five datasets are reported in Table 5. As observed, DEIQT achieves the best performance on four of five datasets and reaches the competing performance on the KonIQ dataset. These results manifest the strong generalization capability of DEIQT.

Data-Efficient Learning Validation

One of the key properties of DEIQT is data-efficient learning, which allows our model to achieve a competing performance to state-of-the-art BIQA methods while requiring substantially less training data. Given the costly image annotation and model training, such a property is highly desired for BIQA methods. To further investigate this property, we conduct controlled experiments to train our model by varying the amount of training data from 20% to 60% with an interval of 20%. We repeat the experiment 10 times for each amount of training data and report the medians of SRCC. The testing data remains 20% images irrespective of the amount of the training data and is completely nonoverlapped with the training data throughout all experiments.

The experimental results are detailed in Tables 6-7. On the synthetic datasets, even with 20% images, DEIQT has reached a competing performance to the second best BIQA method in Table 1. When the training data contains 40% images, DEIQT outperforms all other BIQA methods. In other words, DEIQT utilizes only half of the training data and achieves the best performance on the synthetic datasets. For authentic datasets, DEIQT is capable of achieving the competing performance with 60% images, which is still much more efficient than other BIQA methods.

In addition to the required training data, we further evaluate the training efficiency of DEIQT which is also an important indicator for data-efficient learning. Fig. 5 illustrates the medians of SRCC as the number of epochs increases on the testing set of LIVEC and KonIQ. ViT-BIQA directly utilizes the extracted features of the CLS token to predict the image quality. As shown in Fig. 5, DEIQT converges significantly faster than other methods, where it reaches a satisfactory performance in only two epochs. As a comparison, ViT-BIQA exhibits a slow convergence rate, especially on the small-scale dataset LIVEC. These observations vividly

		LIVE	
Mode	Methods	PLCC	SRCC
20%	ViT-BIQA	0.828	0.894
	HyperNet	0.950	0.951
	DEIQT	**0.968**	**0.965**
40%	ViT-BIQA	0.847	0.903
	HyperNet	0.961	0.959
	DEIQT	**0.973**	**0.971**
60%	ViT-BIQA	0.856	0.915
	HyperNet	0.963	0.960
	DEIQT	**0.974**	**0.972**

Table 6: Data-efficient learning validation with the training set (LIVE) containing 20%, 40% and 60% images. Bold entries indicate the best performance.

		LIVEC		KonIQ	
Mode	Methods	PLCC	SRCC	PLCC	SRCC
20%	ViT-BIQA	0.641	0.622	0.855	0.825
	HyperNet	0.809	0.776	0.873	0.869
	DEIQT	**0.822**	**0.792**	**0.908**	**0.888**
40%	ViT-BIQA	0.714	0.684	0.901	0.880
	HyperNet	0.849	0.832	0.908	0.892
	DEIQT	**0.855**	**0.838**	**0.922**	**0.903**
60%	ViT-BIQA	0.739	0.705	0.916	0.903
	HyperNet	0.862	0.843	0.914	0.901
	DEIQT	**0.877**	**0.848**	**0.931**	**0.914**

Table 7: Data-efficient learning validation with the training set (LIVEC and KonIQ) containing 20%, 40% and 60% images. Bold entries indicate the best performance.

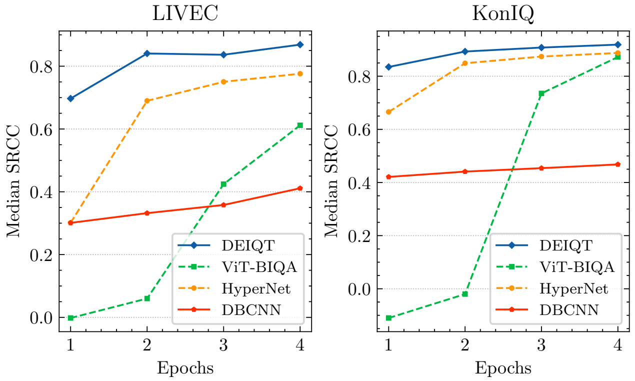

Figure 5: Median SRCC versus Epochs on the LIVEC and KonIQ testing datasets.

		LIVEC		KonIQ	
Module	#Params	PLCC	SRCC	PLCC	SRCC
ViT std	22M	0.770 ±0.045	0.714 ±0.039	0.919 ±0.011	0.908 ±0.011
ViT + AP/6 std	22M	0.782 ±0.033	0.720 ±0.030	0.924 ±0.010	0.913 ±0.008
ViT + Decoder(R*) std	24M	0.871 ±0.018	0.842 ±0.024	0.927 ±0.007	0.916 ±0.006
ViT + Decoder(CLS) std	24M	0.881 ±0.018	0.863 ±0.019	0.931 ±0.005	0.918 ±0.007
DEIQT std	24M	**0.894** ±0.014	**0.875** ±0.017	**0.934** ±0.003	**0.921** ±0.004

Table 8: Ablation experiments on LIVEC and KonIQ datasets. Bold entries indicate the best performance.

demonstrate that DEIQT can efficiently implement the domain adaptation from the pre-training of classification tasks to the fine-tuning of IQA tasks, thereby greatly improving the training efficiency.

Ablation Study

DEIQT is composed of three essential components, including the ViT encoder, quality-aware decoder, and the attention-panel mechanism. We conduct the ablation experiments to examine the individual contribution of each component. Table 8 shows the experimental results on the LIVEC and KonIQ datasets. The ViT in Table 8 refers to the DEIQT without the quality-aware decoder and the attention-panel. It is equivalent to the ViT-BIQA in Fig. 5. AP/6 indicates the attention-panel (AP) with 6 panel members. ViT + AP/6 skips the decoder and sends the inputs of the DEIQT decoder to an MLP head to make the prediction. Decoder(R*) and Decoder(CLS) mean that random numbers or CLS token are utilized as inputs of the decoder, respectively. The proposed DEIQT consists of ViT, Decoder(CLS) and AP/6.

From Table 8, we observe that both the quality-aware decoder and the attention-panel mechanism are highly effective in characterizing the image quality, and thus contributing to the overall performance of DEIQT. In particular, the proposed quality-aware decoder significantly improves the model performance in terms of accuracy and stability, whereas the attention-panel contributes less than the decoder. This is reasonable considering that the number of parameters introduced by the decoder is substantially higher than those introduced by the attention-panel. The operations involved in the decoder are also much more sophisticated. Nevertheless, the attention-panel allows our model to attain improved performance with negligible additional expense.

Finally, we carry out the experiment to investigate the effects of the decoder depth on the DEIQT. The experimental results are listed in Table 9. As can be seen that DEIQT is insensitive to the depth of decoder. When the number of layers of decoder increases, the performance of DEIQT remains almost unchanged on these two datasets. Thus, we set the number of layers of decoder to 1 to maintain a lightweight design for our model.

		LIVEC		KonIQ	
Layer Nums	#Params	PLCC	SRCC	PLCC	SRCC
1	24M	0.894	0.875	0.934	0.921
2	26M	0.890	0.871	0.933	0.919
4	31M	**0.895**	**0.877**	**0.936**	**0.922**
8	40M	**0.895**	0.873	0.933	0.918

Table 9: The effects of the layer numbers of the decoder on the DEIQT. Bold entries indicate the best results.

Conclusion

In this paper, we present a data-efficient image quality transformer (DEIQT), which can accurately characterize the image quality using much less data. In particular, we regard the CLS token as the abstractness of quality-aware features and adapt it to the queries of the dedicatedly designed decoder. Then, we leverage the cross-attention mechanism to decouple the quality-aware features from the encoder outputs. Furthermore, inspired by the human behaviors of the subjective evaluation, we offer a novel attention-panel mechanism to mitigate the prediction uncertainty while introducing almost no additional parameters. Experiments on eight standard datasets demonstrate the superiority of DEIQT in terms of prediction accuracy, training efficiency, and generalization capability.

Acknowledgments

This research was partly supported by the National Key R&D Program of China (Grant No. 2020AAA0108303), and in part by National Natural Science Foundation of China under Grant 62201538, Natural Science Foundation of Shandong Province under grant ZR2022QF006, and in part by China National Postdoctoral Program for Innovative Talents (BX20220392), China Postdoctoral Science Foundation (2022M711729)

References

Bosse, S.; Maniry, D.; Muller, K.-R.; Wiegand, T.; and Samek, W. 2017. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans. Image Process.*, 27(1): 206–219.

Deng, J.; Dong, W.; Socher, R.; Li, L.-J.; Li, K.; and Fei-Fei, L. 2009. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, 248–255.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*. Virtual Event, Austria.

Fang, Y.; Zhu, H.; Zeng, Y.; Ma, K.; and Wang, Z. 2020. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3677–3686.

Ghadiyaram, D.; and Bovik, A. C. 2015. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Trans. Image Process.*, 25(1): 372–387.

Golestaneh, S. A.; Dadsetan, S.; and Kitani, K. M. 2022. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 1220–1230.

Hosu, V.; Lin, H.; Sziranyi, T.; and Saupe, D. 2020. KonIQ-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Trans. Image Process.*, 29: 4041–4056.

Hu, R.; Liu, Y.; Gu, K.; Min, X.; and Zhai, G. 2021a. Toward a No-Reference Quality Metric for Camera-Captured Images. *IEEE Transactions on Cybernetics*.

Hu, R.; Liu, Y.; Wang, Z.; and Li, X. 2021b. Blind quality assessment of night-time image. *Displays*, 69: 102045.

Hu, R.; Monebhurrun, V.; Himeno, R.; Yokota, H.; and Costen, F. 2019. A statistical parsimony method for uncertainty quantification of FDTD computation based on the PCA and ridge regression. *IEEE Transactions on Antennas and Propagation*, 67(7): 4726–4737.

Ke, J.; Wang, Q.; Wang, Y.; Milanfar, P.; and Yang, F. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5148–5157.

Kim, J.; and Lee, S. 2016. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing*, 11(1): 206–220.

Larson, E. C.; and Chandler, D. M. 2010. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1): 011006.

Lin, H.; Hosu, V.; and Saupe, D. 2019. KADID-10k: A large-scale artificially distorted IQA database. In *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*, 1–3. IEEE.

Ma, K.; Liu, W.; Zhang, K.; Duanmu, Z.; Wang, Z.; and Zuo, W. 2017. End-to-end blind image quality assessment using deep neural networks. *IEEE Trans. Image Process.*, 27(3): 1202–1213.

Mittal, A.; Moorthy, A. K.; and Bovik, A. C. 2012. No-reference image quality assessment in the spatial domain. *IEEE Trans. Image Process.*, 21(12): 4695–4708.

Moorthy, A. K.; and Bovik, A. C. 2011. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE Trans. Image Process.*, 20(12): 3350–3364.

Ponomarenko, N.; Jin, L.; Ieremeiev, O.; Lukin, V.; Egiazarian, K.; Astola, J.; Vozel, B.; Chehdi, K.; Carli, M.; Battisti, F.; et al. 2015. Image database TID2013: Peculiarities, results and perspectives. *Signal processing: Image communication*, 30: 57–77.

Raghu, M.; Unterthiner, T.; Kornblith, S.; Zhang, C.; and Dosovitskiy, A. 2021. Do Vision Transformers See Like Convolutional Neural Networks? In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Saad, M. A.; Bovik, A. C.; and Charrier, C. 2012. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE Trans. Image Process.*, 21(8): 3339–3352.

Sheikh, H. R.; Sabir, M. F.; and Bovik, A. C. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans. Image Process.*, 15(11): 3440–3451.

Soundararajan, R.; and Bovik, A. C. 2011. RRED indices: Reduced reference entropic differencing for image quality assessment. *IEEE Trans. Image Process.*, 21(2): 517–526.

Su, S.; Yan, Q.; Zhu, Y.; Zhang, C.; Ge, X.; Sun, J.; and Zhang, Y. 2020. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3667–3676.

Touvron, H.; Cord, M.; Douze, M.; Massa, F.; Sablayrolles, A.; and Jegou, H. 2021a. Training data-efficient image transformers & distillation through attention. In *International Conference on Machine Learning*, volume 139, 10347–10357.

Touvron, H.; Cord, M.; and Jégou, H. 2022. DeiT III: Revenge of the ViT. In *Computer Vision – ECCV 2022: 17th European Conference*, 516–533.

Touvron, H.; Cord, M.; Sablayrolles, A.; Synnaeve, G.; and Jégou, H. 2021b. Going deeper with Image Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 32–42.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, Z.; Bovik, A. C.; Sheikh, H. R.; and Simoncelli, E. P. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Trans. Image Process.*, 13(4): 600–612.

Ying, Z.; Niu, H.; Gupta, P.; Mahajan, D.; Ghadiyaram, D.; and Bovik, A. 2020. From patches to pictures (PaQ-2-PiQ): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3575–3585.

You, J.; and Korhonen, J. 2021. Transformer for image quality assessment. In *2021 IEEE International Conference on Image Processing (ICIP)*, 1389–1393. IEEE.

You, Y.; Li, J.; Reddi, S.; Hseu, J.; Kumar, S.; Bhojanapalli, S.; Song, X.; Demmel, J.; Keutzer, K.; and Hsieh, C.-J. 2020. Large Batch Optimization for Deep Learning: Training BERT in 76 minutes. In *International Conference on Learning Representations*.

Zhang, L.; Zhang, L.; and Bovik, A. C. 2015. A Feature-Enriched Completely Blind Image Quality Evaluator. *IEEE Trans. Image Process.*, 24(8): 2579–2591.

Zhang, W.; Ma, K.; Yan, J.; Deng, D.; and Wang, Z. 2018. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology*, 30(1): 36–47.

Zhang, Y.; Qu, Y.; Xie, Y.; Li, Z.; Zheng, S.; and Li, C. 2021. Perturbed Self-Distillation: Weakly Supervised Large-Scale Point Cloud Semantic Segmentation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15520–15528.

Zhang, Y.; Xie, Y.; Li, C.; Wu, Z.; and Qu, Y. 2022. Learning All-In Collaborative Multiview Binary Representation for Clustering. *IEEE Transactions on Neural Networks and Learning Systems*, 1–14.

Zhu, H.; Li, L.; Wu, J.; Dong, W.; and Shi, G. 2020. MetaIQA: Deep Meta-Learning for No-Reference Image Quality Assessment. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.