



# METER: Multi-task efficient transformer for no-reference image quality assessment

Pengli Zhu<sup>1,2</sup> · Siyuan Liu<sup>1</sup> · Yancheng Liu<sup>1</sup> · Pew-Thian Yap<sup>3</sup>

Accepted: 10 October 2023 / Published online: 6 November 2023  
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2023

## Abstract

No-reference image quality assessment (NR-IQA) is a fundamental yet challenging task in computer vision. Current NR-IQA methods based on convolutional neural networks typically employ deeply-stacked convolutions to learn local features pertinent to image quality, neglecting the importance of non-local information and distortion types. As a remedy, we introduce in this paper an end-to-end multi-task efficient transformer (METER) for the NR-IQA task, consisting of a multi-scale semantic feature extraction (MSFE) backbone module, a distortion type identification (DTI) module, and an adaptive quality prediction (AQP) module. METER identifies the distortion type using the DTI module to facilitate extraction of distortion-specific features via the MSFE module. METER scores image quality in an adaptive manner by adjusting the weights and biases of adaptive fully-connected (AFC) layers in the AQP module, increasing generalizability to images captured in different natural environments. Experimental results demonstrate that METER significantly outperforms existing methods for accuracy and efficiency across five public datasets: LIVEC, BID, KonIQ, LIVE, and CSIQ, and exhibits remarkable performance with Pearson's linear correlation coefficients: 0.923, 0.912, 0.937, 0.978, and 0.982 on respective datasets when compared to human subjective scores. Additionally, METER also attains higher efficiency (-53.9% Params and -87.7% FLOPs) compared to the existing transformer-based methods, making it valuable for real-world applications.

**Keywords** Image quality assessment · Multi-task learning · Efficient transformer

## 1 Introduction

With the growth of social media and the increasing demand for imaging services, an increasingly large amount of image

---

Siyuan Liu contributed equally to this work.

---

✉ Siyuan Liu  
dmu.s.y.liu@gmail.com

Pengli Zhu  
dlmu.p.l.zhu@gmail.com

Yancheng Liu  
liuyc@dlmu.edu.cn

Pew-Thian Yap  
ptyap@med.unc.edu

<sup>1</sup> The College of Marine Engineering, Dalian Maritime University, Dalian 116026, China

<sup>2</sup> The College of Design and Engineering, National University of Singapore, Singapore 119077, Singapore

<sup>3</sup> The Department of Radiology and Biomedical Research Imaging Center (BRIC), University of North Carolina at Chapel Hill, Chapel Hill 27599, USA

data is becoming available to consumers, where image quality refers to the measure of how well an image accurately represents the visual information contained within it, encompassing various factors such as sharpness, clarity, color accuracy, contrast, and overall visual perception. Since image signals are subject to quality degradation at all processing steps, i.e., acquisition, compression, transmission, and display, image quality assessment (IQA) is crucial in determining whether the acquired image data is reliable and whether reacquisition or image enhancement is required.

IQA methods can be categorized as subjective or objective. While the human visual system (HVS) is the ultimate receiver in most visual communication systems, subjective IQA is time-consuming, costly, labor-intensive, and cannot be directly embedded and optimized in automated systems. In contrast, objective IQA automatically predicts visual quality via computer algorithms and is optimizable for timely IQA. Objective IQA has received increasing attention in recent years [1–3].

According to the availability of distortion-free reference images, objective IQA methods can be divided into three categories: (i) Full-reference IQA (FR-IQA) methods [4–6],

which rely on the full access to reference images, can be further classified as spatial domain methods [4, 5] and transform domain methods [6]; (ii) Reduced-reference IQA (RR-IQA) methods, which rely on partial information from reference images [7–9]; and (iii) No-reference IQA (NR-IQA) methods [10–12], the absence of a pristine or original reference image for comparison during the image quality assessment process, which aim to evaluate the quality of a distorted image without having access to the original, unaltered version of the same image.

In various real-world scenarios, obtaining high-quality reference images might be impractical, limiting the applicability of FR-IQA and RR-IQA methods. Therefore, NR-IQA methods are more suitable for practical applications, including image processing in real-time systems, automated image quality evaluation in surveillance, and quality assessment in image compression and transmission pipelines, as they can evaluate image quality using only the distorted image itself without requiring a pristine reference image.

Image quality is directly affected by the type of distortion. NR-IQA methods can be further categorized as distortion-specific or distortion-agnostic. Distortion-specific methods can estimate the perceived quality of a distorted image only if the type of distortion (e.g., JPEG compression [13, 14], blurring, and noise [15, 16]) is known beforehand. They are less generalizable across distortions, leading to IQA failure in real-world applications where distortion types are typically not known in advance. More attention has been recently dedicated to developing distortion-agnostic NR-IQA methods, which are typically based on natural scene statistics (NSS) [17–19] and HVS models [10, 20, 21]. Most NSS methods measure the distribution of image statistics by fitting a NSS model. However, distortion not only changes image statistics, but also disrupts statistical regularity in natural images, leading to mismatch between the fitted NSS model and the actual distribution of image statistics [22]. HVS methods mainly relate the characteristics of the HVS to a mathematical model to mimic human subjective IQA. Although the HVS model can resolve the inconsistency of subjective and objective IQA to some extent, it is limited by high computational complexity and low accuracy.

Recently, learning-based general-purpose models [23–26] have demonstrated great potential in NR-IQA, as they can automatically extract IQA-pertinent features and regress the quality score in an end-to-end fashion. While convolutional neural networks (CNNs) [1, 27, 28] have demonstrated promising performance, they suffer from three main drawbacks: 1) Most CNNs learn local but ignore non-local distortion features, leading to quality prediction that is inconsistent with the HVS; 2) Most CNNs are inconsistent with the perceptual order by which the HVS evaluates image quality, i.e., understanding image content (e.g., objects in a scene) before recognizing whole image information; and 3) Unlike

subjective IQA, information on the type of distortion is typically neglected in most CNN-based methods, leading to poor IQA performance and low generalizability.

Unlike CNNs that have a strong bias towards learning local features, the transformer [29–31] is effective for learning long-range dependencies. The ability to capture pairwise interactions enhances the sensitivity of the transformer to local distortions and global contents. Transformers can therefore mimic HVS for accurate IQA in a holistic manner. Despite its great potential, the transformer has high memory and computational costs, since it scales quadratically with spatial or embedding dimensions, resulting in vast overheads in training and inference.

In this paper, we draw inspiration from previous work [1, 2, 32] and propose the end-to-end multi-task efficient transformer (METER) for NR-IQA, predicting simultaneously the distortion type and the quality score. The key features of METER are as follows:

- (i) To improve the accuracy and efficiency of NR-IQA, we propose a novel method called METER, which is based on an efficient multi-task transformer and consists of three main components: a multi-scale semantic feature extraction (MSFE) module, which extracts both global and local features at multiple scales; an adaptive quality prediction (AQP) module, which performs IQA using adaptive fully connected (AFC) layers; and a distortion identification (DTI) module, which predicts image distortion type to facilitate the extraction of distortion-specific IQA-pertinent features. METER extracts and aggregates local and global features, and incorporates the identified distortion type as an additional input for quality regression.
- (ii) To efficiently extract multi-scale features for better quality prediction, we introduce the MSFE module based on efficient transformer blocks (ETBs) in four stages. In each stage, we employ patch embedding for resolution reduction and channel dimension extension, position encoding, and ETB for self-attention feature extraction. We also introduce a local distortion aware (LDA) modules to capture multi-scale local features from outputs of different stages, which are then assembled with global features to construct multi-scale IQA-pertinent representations for subsequent distortion type identification and quality regression.
- (iii) To mimic the perceptual model of HVS, we propose AFC layers, whose hyperparameters, i.e., weight and bias, are learned from global features for content-aware AQP. Since distortion type is an important factor in determining image quality, we employ a multi-task framework to merge AQP with distortion-type probabilities predicted by the DTI module for further improvement of IQA accuracy.

The remainder of this paper is organized as follows. In Section 2, we briefly review related work on multi-task IQA and transformers for IQA, as well as the efficient visual transformer. In Section 3, we describe the overall architecture of METER and module details. In Section 4, we evaluate METER in comparison with state-of-the-art methods and perform ablation studies. In Section 5, we discuss the relationship between METER and existing methods and analyze the main limitations. Finally, we conclude in Section 6.

## 2 Related work

### 2.1 Multi-task IQA

In visual communication systems, all imaging steps, i.e., acquisition, processing, compression, transmission, and decompression, can introduce distortions that can have different effects on image quality. In comparison with single-task IQA methods [33–35], multi-task IQA has a wider range of functionality and usefulness. Most multi-task IQA methods can be classified into two categories: sequential methods [1, 36], which assess image quality by first determining the distortion type, and parallel methods [3, 27, 37], which determine the distortion type and assess the image quality simultaneously. Our work is inspired by MEON [1], a parallel multi-task method that incorporates distortion prediction into an IQA network to guide quality prediction. In addition, the types of distortions that METER specifically tackles in this paper are JPEG-2000 Compression (JP2K), JPEG Compression (JPEG), White noise (WN), Gaussian blurring (GB), Fast Fading Rayleigh (FF), additive white Gaussian noise (AWGN), global contrast decrements (CC) and additive pink Gaussian noise (FN), which are mainly present in the LIVE and CSIQ datasets.

### 2.2 Transformers for IQA

CNNs are typically used as the main backbone for feature extraction in the majority of IQA models [2, 28, 38], but they have a strong locality bias and is ineffective in capturing non-local information.

Transformers, introduced by Vaswani et al. [29], leverage self-attention mechanisms [39] to aggregate information from an entire data sequence, making them effective for natural language processing and computer vision tasks [40–42]. However, the application of transformer to NR-IQA is not fully explored. Concurrent with our work, transformers were used for NR-IQA [43] by feeding features from the last layer of a CNN to a transformer for quality prediction. A hybrid approach taking advantage of a CNN and the self-attention mechanism in a transformer was leveraged to extract both local and non-local features from the input image

for NR-IQA [30]. Unlike these methods [30, 31, 44], we use the transformer to learn the mapping between an input image and multi-scale semantic features for adaptive NR-IQA that mimics the HVS.

### 2.3 Efficient visual transformer

To model long-range contextual information, self-attention [29, 45, 46] has been introduced in transformers for visual recognition tasks to learn dependencies across channels or in space by appending self-attention layers atop convolutional networks. However, the nature of self-attention in transformers causes greater computational overhead over conventional CNNs. In view of this, efficient variants have been proposed, such as Unified transFormer (UniFormer) [47] and flexible transformer (FlexFormer) [48], aiming to trade off between accuracy and efficiency in optimizing transformers for real-world applications. This allows lightweight efficient transformers to be employed for image super-resolution [49, 50], semantic segmentation [51, 52], and object detection [53, 54]. In this work, we draw inspiration from an efficient multi-scale vision transformer, called ResT [32], for multi-scale, memory-efficient, and multi-head self-attention.

## 3 Method architecture

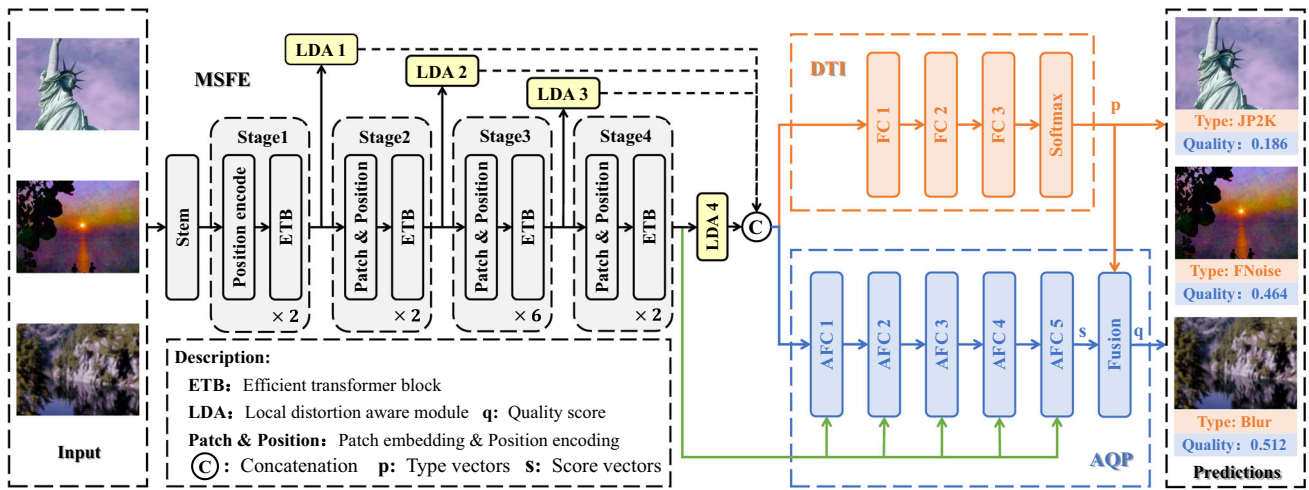
Figure 1 shows an overview of METER, which consists of a multi-scale semantic feature extraction (MSFE) module, a distortion-type identification (DTI) module, and an adaptive quality prediction (AQP) module. The MSFE module extracts multi-level local and global features, which are then merged into two mutually collaborative modules, i.e., AQP and DTI. The DTI module encodes degradation-style features and predicts the distortion type. The AQP module incorporates the predicted distortion type for final QA by adaptively learning weights and biases of adaptive fully-connected (AFC) layers. Details on the METER will be described next.

### 3.1 Overview of meter

Deep learning-based IQA models typically learn a nonlinear mapping  $f$  between an input image  $x$  and its quality score  $q$ :

$$q = f(x). \quad (1)$$

To mimic human visual perception, we extract multi-scale features and identify the distortion type to guide IQA in a multi-task framework. We extract multi-scale semantic features  $R = [r^1, \dots, r^m] = f_r(x)$  via a nonlinear mapping  $f_r$  learned with the MSFE module in each stage, where  $m = 4$ . Based on  $R$ , the DTI module classifies the distortion into



**Fig. 1** Overview of METER. Given an image, we first extract multi-level local and global features using a multi-scale semantic feature extraction (MSFE) module and merge them in two mutually collaborative modules, i.e., a distortion-type identification (DTI) module and an adaptive quality prediction (AQP) module. The DTI module encodes

degradation-style features and predicts the distortion type, and the AQP module incorporates the predicted distortion type for final quality prediction by adaptively learning weights and biases of adaptive fully-connected (AFC) layers

multiple types, giving a distortion-type probably vector  $p$ . To adapt to different image contents and distortion types, AQP embeds  $R$  using AFC layers, with weights and biases dynamically learned from global features  $r^m$ , and incorporates  $p$  to regress the image quality score, i.e.,  $q = f_q(R, p)$ . The objective of METER is therefore:

$$\begin{aligned} \min L &= \min \sum_{i=1}^n L_i \{q_i, q_i^*\} \\ &= \min \sum_{i=1}^n L_i \{f_q(R_i, p_i), q_i^*\}, \end{aligned} \tag{2}$$

where  $n$  is the batch size,  $r_i^m$  is the  $m$ -th element of  $R_i = [r_i^1, \dots, r_i^m]$  obtained from the  $x_i$  via the MSFE module,  $x_i$  is the  $i$ -th image in the batch,  $q_i^*$  is ground-truth quality score of  $x_i$ . The loss  $L_i$  measures the difference between  $q_i$  and  $q_i^*$ . Next, three modules of METER are described next.

### 3.2 Multi-scale Semantic Feature Extraction (MSFE) backbone module

The MSFE backbone module (Fig. 1) consists of two sub-modules: 1) an efficient vision transformer (EVT, gray in Fig. 1), consisting of a stem module applied for low-level information extraction, followed by four stages to capture multi-scale feature maps. Each stage consists of a patch embedding block, a position encoding block, and an efficient transformer blocks (ETBs); and 2) four local distortion-aware (LDA, yellow in Fig. 1) blocks for multi-scale feature

alignment (Fig. 3), where a  $1 \times 1$  convolution followed by an averaging pooling (AvgPool) layer and a fully connected layer are utilized to construct LDA blocks 1 to 4. LDA block 4 omits the  $1 \times 1$  convolution.

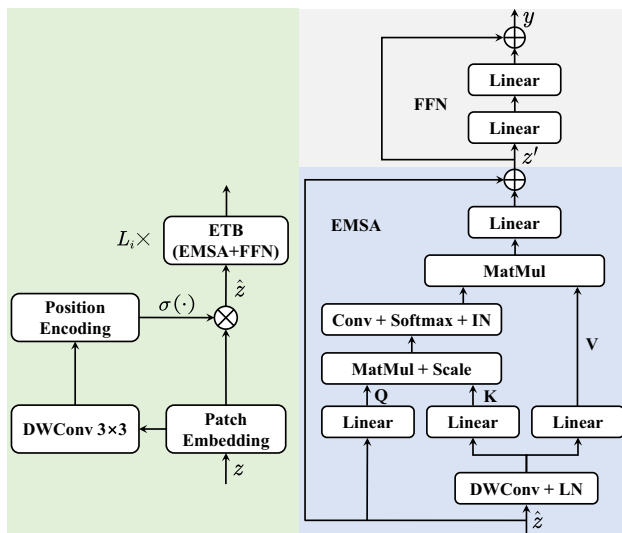
*Efficient Vision Transformer (EVT):* To effectively extract low-level features, we first adopt a stem module that consists of two  $3 \times 3$  standard convolution layers with respectively strides 2 and 1 followed by batch normalization [55], ReLU activation [56], and a  $3 \times 3$  standard convolution layer with stride 2. We consider the extracted low-level features as a sequence and then introduce four stages to extract local and global semantic features.

Specifically, in Stage 1, we first utilize a simple and effective pixel attention (PA) module for position encoding, applying a  $3 \times 3$  depth-wise convolution (DWConv) and a sigmoid activation  $\sigma(\cdot)$  [57] to calculate pixel-wise weights, which are combined with the encoded input  $z$  to generate position encoding features  $\hat{z}$ :

$$\hat{z} = \text{PA}(z) = z * \sigma(\text{DWConv}(z)). \tag{3}$$

We then take the position encoding features as input tokens and encode them using efficient transformer blocks (ETBs), as illustrated in Fig. 2. Each ETB consists of 1) efficient multi-head self-attention (EMSA, blue in Fig. 2), which employs query (Q) - key (K) - value (V) decomposition to model global relationships between sequence tokens; and 2) an efficient feed-forward network (FFN, gray in Fig. 2), which employs a residual framework to transform features and learn more complete representations.





**Fig. 2** Overall architecture of each stage in Fig. 1. (Green) We take the position encoding features as input tokens and encode them using efficient transformer blocks (ETBs), which consists of two parts: (Blue) Efficient multi-head self-attention (EMSA), which employs a query (**Q**) - key (**K**) - value (**V**) decomposition to model global relationships between sequence tokens, and (Gray) feed-forward network (FFN), which employs a residual framework to transform features and learn wider representations

In Stages 2–4, patch embedding implemented via a  $3 \times 3$  standard convolution and batch normalization (BN) [55] are employed before position encoding to reduce the spatial dimension by  $4 \times$  and expand the channel dimension by  $2 \times$ .

For EMSA, the 2D input token  $\hat{z} \in \mathbb{R}$  is first linearly projected to query **Q** using a fully-connected layer. Simultaneously, the 2D input token is reshaped to 3D along the spatial dimension to compress computational memory and then employ a DWConv to further reduce the height and width dimensions by a factor of  $\alpha$ , which is adaptively set by the number of EMSA head  $k$ , i.e.,  $\alpha = 8/k$ . Accordingly, the kernel size, stride, and padding are set to  $\alpha + 1$ ,  $\alpha$ , and  $\alpha/2$ , respectively. Then the key **K** and value **V** of self-attention can be obtained from a linear projection. The self-attention function of **Q**, **K** and **V** are computed as follows:

$$\text{EMSA}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{IN} \left( \text{Softmax} \left( \text{Conv} \left( \frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}} \right) \right) \right) \mathbf{V}, \tag{4}$$

where  $\text{Conv}(\cdot)$  denotes the standard  $1 \times 1$  convolution,  $\text{Softmax}(\cdot)$  is the softmax activation [58], and  $d_k$  is the output dimension. Instance normalization [59],  $\text{IN}(\cdot)$ , is employed on the dot product matrix to maintain the diversity ability of multi-heads. The outputs of all heads are then concatenated and linearly projected to construct the final output of EMSA. In addition, EMSA is followed by FFN and the output for

each ETB can be obtained as

$$y = z' + \text{FFN}(\text{LN}(z')), \tag{5}$$

where  $z' = \hat{z} + \text{EMSA}(\text{LN}(\hat{z}))$ . Layer normalization (LN) [60] is applied before EMSA and FFN to ensure the stability of the data feature distribution for training.

Unlike the traditional MSA that is computationally intensive, EMSA compresses the memory by a simple depth-wise convolution operation. In addition, we compensate short-length limitations of the input token for each head by projecting the interaction across the dimension of attention-heads while keeping the diversity ability of multi-heads.

*Local Distortion-Aware (LDA) block:* To incorporate multi-scale features extracted by EVT, we introduce an LDA module to encode distortion-aware features from local features at each level, as illustrated in Fig. 3. Each LDA is constructed by a fully connected layer, an average pooling layer, and a  $1 \times 1$  standard convolution layer. LDA 4 is without the convolution layer. We take the extracted local semantic features at each stage as input to the corresponding LDA, and then merge the outputs of LDAs,  $r_i^m$  ( $m \in [1, 4]$ ), to obtain the multi-scale semantic features  $R_i = [r_i^1, \dots, r_i^m]$  of  $x_i$ .

### 3.3 Distortion Type Identification (DTI) module

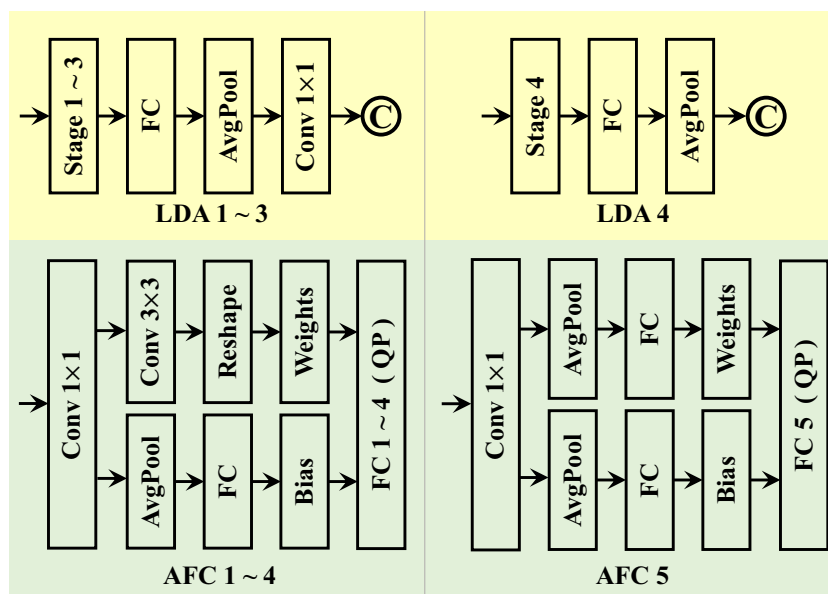
To identify the distortion type for enhancing IQA (orange-pink in Fig. 1), we introduce a distortion type identification module (DTI) to predict the distortion-type probability vector, which indicates the likelihood of each distortion type and will be a partial input to AQP (see Section 3.4). With the multi-scale semantic features  $R_i$  extracted from MSFE as input, DTI utilizes three fully connected layers with ReLU activation to encode multi-scale semantic features to distortion-type pertinent features and employs dropout [61] to prevent overfitting. The final probability vector of distortion types  $p_i$  is obtained via a softmax function:

$$p_i^j(x_i) = \frac{\exp(f_d(f_r(x_i)))}{\sum_{j=1}^C \exp(f_d(f_r(x_i)))}, \quad j \in [1, C], \tag{6}$$

where  $f_r$  and  $f_d$  are the nonlinear mappings of MSFE and DTI, respectively.  $p_i = [p_i^1, \dots, p_i^C]^T$  is a  $C$ -dimensional probability vector of the  $i$ -th input in a mini-batch, indicating the probability of each distortion type, and is fed to AQP to predict the overall perceptual quality score.

### 3.4 Adaptive Quality Prediction (AQP) module

For accurate IQA of images with various contents and different distortion types, we propose an adaptive quality



**Fig. 3** Architectures of LDA and AFC. (Yellow) An LDA module to encode distortion-aware features from local features at each level, where each LDA is constructed with a fully connected layer, an average pooling layer, and a  $1 \times 1$  standard convolution layer (except LDA 4). (Green) An adaptive fully connected (AFC) layer, with weight and bias hyperparameters learned from the global semantic features of the MSFE module.

Each AFC layer utilizes two  $1 \times 1$  standard convolutions with ReLU, a  $3 \times 3$  standard convolution with reshape or fully connected layer to respectively learn its weight matrix and bias vector from global semantic features (except AFC layer 5). The weight and bias of AFC layer 5 are learned by an average pooling layer and fully connected layers

prediction (AQP) module, which is constructed by five adaptive fully-connected layers, and a merging layer for distortion-type probability and image quality regressor, as illustrated in Fig. 1.

To adapt to various image contents, we propose an adaptive fully connected (AFC) layer (Fig. 3), with weight and bias hyperparameters learned from the global semantic features of the MSFE module. Each AFC layer utilizes two  $1 \times 1$  standard convolutions with ReLU, a  $3 \times 3$  standard convolution with reshape or fully connected layer to respectively learn its weight matrix and bias vector from global semantic features (except AFC layer 5). The weight and bias of AFC layer 5 are learned by an average pooling layer and fully connected layers. The output channels of the convolution and fully connected layers used for AFC are determined according to the dimensions of corresponding layers for size matching. With five AFC layers followed by a sigmoid activation, AQP module produces a scoring vector  $s_i$ .

In the fusion layer, with the predicted distortion-type probability vector  $\mathbf{p}_i$  and scoring vector  $\mathbf{s}_i$  as input, the overall perceptual quality score  $q_i$  for  $x_i$  is predicted using probability-weighted summation [62]:

$$q_i = \mathbf{p}_i^\top \mathbf{s}_i = \sum_{i=1}^C p_i \cdot s_i. \tag{7}$$

## 4 Experiments

### 4.1 Experimental settings

#### 4.1.1 Datasets

As summarized in Table 1, we used three realistically distorted image datasets, including LIVE Challenge (LIVEC) [63], KonIQ [64], and BID [65] for evaluation. LIVEC contains 1162 real-world images taken by different photographers with various cameras, covering complex distortions. KonIQ consists of 10073 images selected from the large public multimedia dataset YFCC100m [66], covering a wide range of quality in terms of brightness, colorfulness, contrast, and sharpness. BID contains 586 images with realistic blur distortions such as motion and out-of-focus blur.

**Table 1** Summary of IQA datasets

Dataset	Distorted Img. No.	Distortion Types No.	Source
LIVE	799	5	Synthetic
CSIQ	866	6	Synthetic
BID	590	1	Authentic
LIVEC	1162	–	Authentic
KonIQ	10073	–	Authentic

Apart from authentic image datasets, we also tested our model on synthetic image datasets LIVE [67] and CSIQ [68], which respectively contain 779 and 866 synthetically distorted images, as summarized in Table 1. Among them, LIVE has five different distortion types, i.e., JPEG-2000 Compression (JP2K), JPEG Compression (JPEG), White noise (WN), Gaussian blurring (GB), Fast Fading Rayleigh (FF) and CSIQ has six different distortion types, i.e., JPEG-2000 Compression (JP2K), JPEG Compression (JPEG), additive white Gaussian noise (AWGN), Gaussian blurring (GB), global contrast decrements (CC), additive pink Gaussian noise (FN). For each dataset, we randomly split them into a training set and a test set according to the ratio of 8 : 2.

#### 4.1.2 Compared methods

To demonstrate the effectiveness of METER, ten state-of-the-art NR-IQA methods from three categories were employed for comparison:

- Handcrafted feature methods, i.e., integrated local natural image quality evaluator (ILNIQE) [33], and high order statistics aggregation (HOSA) [34];
- Deep learning synthetic IQA methods, i.e., blind image evaluator based on a convolutional neural network (BIECON) [35], and weighted average deep image quality measure (WaDIQaM) [69];
- CNN-based authentic IQA methods, i.e., semantic feature aggregation (SFA) [28], probabilistic quality representation (PQR) [38], deep bilinear convolutional neural network (DBCNN) [27], hyper-network IQA (Hyper-IQA) [2], distortion graph representation learning for IQA (GraphIQA) [3].
- Transformer-based authentic IQA methods, i.e., transformers, relative ranking, and self consistency (TReS) [30], swin transformer IQA (SwinIQA) [44], and multi-dimension attention network image quality assessment (MANIQA) [31]. MANIQA was ranked first in the final private test phase in the NTIRE 2022 Perceptual IQA Challenge Track 2: No-reference.

To verify the effectiveness of the LDA, DTI and AQP blocks in METER, we compared with 5 ablated versions of METER:

- EVT+DTI: The global features are extracted using EVT, and then the distortion types are identified by DTI;
- EVT+LDA+DTI: LDA block added to EVT+DTI to extract multi-scale features;
- EVT+AQP: The global features are extracted by EVT, and then combined with AQP to dynamically learn image content features and perform quality prediction;

- EVT+LDA+AQP: LDA block added to EVT+AQP for multi-scale feature extraction;
- EVT+LDA+QP+DTI: QP added to EVT+LDA+DTI to perform quality prediction, and the QP module is composed with general FC layers without adaptive prediction capability.

Note that METER is a EVT+LDA+AQP+DTI network. The ablated networks were trained in a fashion similar to the METER.

#### 4.1.3 Evaluation metrics

We adopted two commonly used metrics [70], i.e., Spearman's rank order correlation coefficient (SROCC) and Pearson's linear correlation coefficient (PLCC), to measure the monotonicity and accuracy of IQA results. Both of them range between 0 and 1, and a higher value indicates better performance. Before calculating PLCC, logistic regression is first applied to remove nonlinear effects on the rating caused by human visual observation, as suggested in the report from the Video Quality Expert Group (VQEG) [70]. These metrics are to measure the correlation between a set of estimated visual quality scores  $Q_{est}$  and a set of human subjective quality scores  $Q_{sub}$ :

$$SROCC(Q_{est}, Q_{sub}) = 1 - \frac{6 \sum d_i^2}{N(N^2 - 1)}, \quad (8)$$

$$PLCC(Q_{est}, Q_{sub}) = \frac{cov(Q_{sub}, Q_{est})}{\sigma(Q_{sub})}, \sigma(Q_{est}) \quad (9)$$

where  $N$  is the number of images in the evaluation dataset,  $d_i$  is the rank difference of  $i$ -th evaluation sample in  $Q_{est}$  and  $Q_{sub}$ ,  $cov(\cdot)$  represents the covariance between  $Q_{est}$  and  $Q_{sub}$ ,  $\sigma(\cdot)$  represents the standard deviation. In addition, to evaluate the accuracy of DTI in classifying distortion types, we introduce a distortion classification evaluation index (ACC), which is expressed as

$$ACC = \frac{N_{correct}}{N_{total}} \times 100\%, \quad (10)$$

where  $N_{correct}$  is the number of correct prediction distortion types,  $N_{total}$  is the total number of predicted distortion types. All the results are tested on datasets with 10 random splittings, and the average results (SROCC and PLCC values) are reported. Furthermore, we introduce three efficiency evaluation metrics, namely Parameters (Params), Floating-point Operations (FLOPs), and Frames Per Second (FPS), to assess the efficiency performance of different methods. Lower Params and FLOPs are preferred, indicating reduced computational and memory requirements, while a higher FPS is favored, reflecting better real-time performance.

#### 4.1.4 Implementation details

METER was implemented using Pytorch [71]. Training and testing of METER were based on a machine with a CPU (Intel Xeon Silver 4210) and three GPUs (NVIDIA RTX A6000 48GB). Following the training strategy from [2], we randomly sample and horizontally flip 25 patches with size  $224 \times 224$  pixels from each training image for augmentation. The training patches inherit quality scores from the source image. We minimize the multi-task loss over the training set:

$$L(p_i, q_i) = \frac{1}{N} \sum_i [\lambda_q L_q(q_i, q_i^*) + \lambda_p L_p(p_i, p_i^*)], \quad (11)$$

where the distortion-type classification loss  $L_p$  and quality regression loss  $L_q$  are cross entropy loss and smooth  $L1$  loss respectively,  $i$  is the index in a mini-batch and  $p_i$  is the predicted probability of distortion type,  $q_i$  is the predicted image quality score, and  $p_i^*$  and  $q_i^*$  are ground-truth labels of distortion type and quality assessment respectively. The hyperparameters  $\lambda_p$  and  $\lambda_q$  were set to balance the performance of distortion type classification and image quality regression. For training models on authentic image datasets, we set  $\lambda_p = 0$  due to the lack of annotations of distortion types.

We used Adam [72] optimizer with weight decay  $5 \times 10^{-4}$  to train our model for 15 epochs, with mini-batch size of 384. Learning rate was first set to  $1 \times 10^{-3}$  for the DTI module,  $2 \times 10^{-5}$  for the other modules, and reduced by 10 after every 5 epochs. For faster convergence, the pretrained model on ImageNet [73] was used for MSFE, and the un-pretrained layers of our model were initialized using Xavier [74]. During testing phase, 25 patches with  $224 \times 224$  pixels from testing image were randomly sampled and their corresponding predictions were average-pooled to arrive at the final quality.

## 4.2 Comparison with the state-of-the-art methods

We compared METER with state-of-the-art (SOTA) NR-IQA methods, including hand-craft feature based methods [33, 34], deep learning based synthetic IQA methods [35, 69] and deep learning based authentic IQA methods [2, 3, 27, 28, 30, 31, 38]. All experiments were conducted 10 times to avoid bias.

### 4.2.1 Performance comparison

Table 2 shows the evaluation results in terms of SROCC and PLCC of the testing underwater images, using different UIE methods, i.e., BRISQUE, ILNIEQE, HOSA, BIECON, WaDIQaM, SFA, PQR, DBCNN, HyperIQA, GraphIQA,

TReS, MANIQA, METER without AFC and METER, where bold entries in black and blue are the best and second-best in terms of performance, respectively. From Table 2, METER outperforms all the SOTA methods for both SROCC and PLCC on all authentic image datasets except KonIQ, on which it achieves second-best performance of PLCC evaluations. This suggests that learning image multi-scale and long-range features first assists in perceiving image quality, when image data covers a wide range of variety. Although there is no modules for synthetic image feature extraction, our method still achieves competing or superior performance to the SOTA methods on two synthetic image datasets LIVE and CSIQ. Compared with the SOTA methods, METER with normal FC instead of AFC (denoted as Ours (w/o AFC) in Table 2) can still perform comparably on most datasets.

We further conducted performance comparison of different methods on individual distortion types. Since the distortion types of images are unavailable in authentic datasets, we evaluated the performance on synthetic image datasets, i.e., LIVE and CSIQ, as summarized in Table 3. Likewise, bold entries in black and blue are the best and second-best performances, respectively. Compared with other methods that introduce specific modules to handle synthetic IQA task, such as ILNIQE [33], HOSA [34], BIECON [35], WaDIQaM [69], PQR [38], DBCNN [27], METER outperforms the other compared methods on all distortion types except JPEG type in LIVE and GB type in CSIQ, but the performance difference is quite slight. This demonstrates that METER is not sensitive to the effects of individual distortions.

### 4.2.2 Generalizability comparison

To verify the generalizability of METER, we conducted cross-dataset performance comparisons on intra either authentic or synthetic datasets. We selected the most competing three methods in above experiments, i.e., DBCNN, HyperIQA and TReS for comparison, and the quantitative results are summarized in Table 4, where bold entries in black are the best in terms of performance. From the testings across six authentic datasets (i.e., LIVEC-BID, LIVEC-KonIQ, BID-LIVEC, BID-KonIQ, KonIQ-LIVEC and KonIQ-BID) in Table 4, our method METER achieves four times of the top performance of compared methods. For the testings across synthetic datasets, i.e., LIVE-CSIQ and CSIQ-LIVE, our method METER still performs competitively to other algorithms, indicating the strong generalizability of METER.

### 4.2.3 Efficiency comparison

To verify the efficiency of METER, we performed a efficiency comparison of different methods on images with the same resolution. In order to ensure the fairness, we selected three recently transformer-based IQA methods, i.e., TReS



**Table 2** Performance comparison on different public datasets

Method	SROCC					PLCC				
	LIVEC	BID	KonIQ	LIVE	CSIQ	LIVEC	BID	KonIQ	LIVE	CSIQ
ILNIQE [33]	0.432	0.516	0.507	0.902	0.806	0.508	0.554	0.523	0.865	0.808
HOSA [34]	0.640	0.721	0.671	0.946	0.741	0.678	0.736	0.694	0.947	0.823
BIECON [35]	0.595	0.539	0.618	0.961	0.815	0.613	0.576	0.651	0.962	0.823
WaDIQaM [69]	0.671	0.725	0.797	0.954	0.955	0.68	0.742	0.805	0.963	0.973
SFA [28]	0.812	0.826	0.856	0.883	0.796	0.833	0.840	0.872	0.895	0.818
PQR [38]	0.857	0.775	0.880	0.965	0.873	0.882	0.794	0.884	0.971	0.901
DBCNN [27]	0.851	0.845	0.875	0.968	0.946	0.869	0.859	0.884	0.971	0.959
HyperIQA [2]	0.859	0.869	0.906	0.962	0.923	0.882	0.878	0.917	0.966	0.942
GraphIQA [3]	0.863	0.837	0.907	0.976	0.943	0.886	0.844	0.922	0.976	0.956
TReS [30]	0.851	0.858	0.915	0.968	0.922	0.871	0.868	0.928	0.968	0.942
MANIQA [31]	<b>0.885</b>	<b>0.896</b>	<b>0.923</b>	<b>0.982</b>	<b>0.968</b>	<b>0.916</b>	<b>0.897</b>	<b>0.942</b>	<b>0.983</b>	0.961
METER (w/o AFC)	0.849	0.864	0.912	0.965	0.966	0.890	0.882	0.925	0.967	<b>0.968</b>
METER	<b>0.896</b>	<b>0.899</b>	<b>0.929</b>	<b>0.977</b>	<b>0.982</b>	<b>0.923</b>	<b>0.912</b>	<b>0.937</b>	<b>0.978</b>	<b>0.982</b>

[30], SwinIQA [44], and MANIQA [31] for comparison, and the quantitative results are summarized in Table 5. In the efficiency comparison experiment, all models are evaluated on  $224 \times 224$  resolution, FPS (*images/s*) is measured on a single RTX A6000 GPU. From Table 5, METER outperforms all the comparison methods for Params, FLOPs and FPS, achieves the best performance of efficiency evaluations. This suggests that METER has good efficiency while maintaining accuracy, which is valuable for real-world scenario applications.

### 4.3 Interpretation and visualization

In order to validate the effectiveness of METER more intuitively, we interpret and visualize the MSFE features and

AFC weights to further explain the role of the different modules in the METER.

#### 4.3.1 Interpretation of MSFE features

We randomly sampled 10 images from the KonIQ [64], and use group score-weighted Class Activation Mapping (Group-CAM) [75] to visualize the attentions at the last stage of the MSFE using heatmaps, as shown in Fig. 4. It can be observed from Fig. 4 that the proposed MSFE can adaptively produce attention maps according to the image contents. This demonstrates that METER understands the image content (e.g., objects in a scene) before recognizing global image information, which is more consistent with the perceptual order by which the HVS evaluates image quality.

**Table 3** SROCC comparisons of various distortion types on the LIVE and CSIQ datasets

Dataset Type	LIVE					CSIQ					
	JP2K	JPEG	WN	GB	FF	JP2K	JPEG	AWGN	GB	CC	FN
ILNIQE [33]	0.894	0.941	0.981	0.915	0.833	0.906	0.899	0.850	0.858	0.501	0.874
HOSA [34]	0.935	0.954	0.975	0.954	0.954	0.818	0.733	0.604	0.841	0.716	0.500
BIECON [35]	0.952	<b>0.974</b>	0.980	0.956	0.923	0.954	0.942	0.902	0.946	0.523	0.884
WaDIQaM [69]	0.942	0.953	0.982	0.938	0.923	0.947	0.853	<b>0.974</b>	<b>0.979</b>	0.923	0.882
PQR [38]	0.953	0.965	0.981	0.944	0.921	0.955	0.934	0.915	0.921	0.837	0.926
DBCNN [27]	0.955	0.972	0.980	0.935	0.930	0.953	0.940	0.948	0.947	0.870	0.940
HyperIQA [2]	0.949	0.961	0.982	0.926	0.936	<b>0.960</b>	0.934	0.927	0.915	0.874	0.931
GraphIQA [3]	0.965	0.966	<b>0.984</b>	0.930	0.954	0.939	0.921	0.939	0.947	0.927	0.919
TReS [30]	0.962	0.904	0.923	0.937	<b>0.960</b>	0.948	0.931	0.940	0.942	0.832	0.937
MANIQA [31]	<b>0.967</b>	<b>0.979</b>	0.983	<b>0.986</b>	0.956	0.953	<b>0.958</b>	0.941	0.959	<b>0.928</b>	<b>0.985</b>
METER	<b>0.978</b>	0.972	<b>0.997</b>	<b>0.989</b>	<b>0.990</b>	<b>0.980</b>	<b>0.985</b>	<b>0.985</b>	<b>0.978</b>	<b>0.945</b>	<b>0.987</b>

**Table 4** Generalizability comparison of cross-dataset testing on SROCC

Training	Testing	DBCNN	HyperIQA	TReS	METER
LIVEC	BID	0.762	0.756	0.713	<b>0.858</b>
	KonIQ	0.754	0.772	0.734	<b>0.787</b>
BID	LIVEC	0.725	0.770	<b>0.779</b>	0.748
	KonIQ	<b>0.724</b>	0.688	0.671	0.687
KonIQ	LIVEC	0.755	0.785	0.786	<b>0.798</b>
	BID	0.816	0.819	0.825	<b>0.833</b>
LIVE	CSIQ	0.758	0.744	0.761	<b>0.769</b>
CSIQ	LIVE	0.877	<b>0.926</b>	0.884	0.901

### 4.3.2 Visualization of AFC hyperparameters

To verify the effectiveness of the weight generation process in AFCs, we extracted the weights and biases of the first AFC in AQP module for several image pairs that are with the same background and object type but different object appearances, e.g., shapes and dressings, and then used PCA transformation [76] to plot them in 3D space for visualization in Fig. 5. From Fig. 5, it is observed that each image can be located by a weight and a bias, and weights and biases differ with image pairs. Also, image pairs generate near spatial distances in a horizontal plane, which indicate similar adaptive weights of AFCs, and the spatial distances of weights for different pair of images are obvious to distinguish. Besides, the corresponding biases are also distinguished from each other for image pairs, such as the planes, animals, plants and athletes in Fig. 5. The analysis above shows that METER understands high-level image content features and avoid image quality errors due to changes of image contents.

### 4.3.3 Visualization result of each dataset

We performed quality prediction for all images in the five datasets and visualized the best and worst results in each dataset, as shown in Fig. 6. The best quality prediction results have clearer content features, and even though the images in LIVE and CSIQ are affected by JPEG and FN distortions,

**Table 5** Efficiency comparisons with different transformer-based IQA methods

Method	Resolution	Params ↓	FLOPs ↓	FPS ↑
TReS [30]	224 × 224	152.45M	16.8G	50.44
SwinIQA [44]	224 × 224	30.81M	51.35G	40.72
MANIQA [31]	224 × 224	127.74M	108.61G	55.44
METER	224 × 224	14.18M	2.06G	73.05

METER can still discriminate the distortion amount based on the DTI module to ensure the accuracy of quality prediction. As for the worst results, the content information of the images is not clear. Specifically, the image in LIVE suffers from chromatic aberration and deformation due to FF distortion, which affects the semantic feature extraction of METER. In addition, the images of CSIQ and LIVEC are so blurred that they cannot effectively adjust the quality scores based on the content information, and the image features of BID and KonIQ are dominated by irregular landscapes, which sometimes causes the quality prediction logic of METER to be inconsistent with HVS, and thus results in predicted scores that differ significantly from the labels.

### 4.4 Ablation study

To verify the effectiveness of LDA, DTI and AQP in METER, we conduct several ablation experiments on the LIVEC, LIVE and CSIQ datasets. We first use a pretrained MSFE on Imagenet [73] as our backbone model and analyze the effect of each individual component by comparing SROCCs, PLCCs and ACCs. The results are shown in Table 6, where bold entries in black are the best in terms of performance.

*Effectiveness of LDA module:* Compared with EVT+DTI and EVT+LDA+DTI, the LDA block improves the ACC of distortion-type classification on the LIVE and CSIQ datasets by over 1.2% and 1.4%, respectively. Compared with EVT+AQP and EVT+LDA+AQP, the LDA block improves the SROCC and PLCC of IQA on the LIVEC, LIVE and CSIQ datasets by over 1.2%, 2.1%, 1.0% and 1.2%, 1.1%, 0.7%, respectively. This indicates that the multi-scale features extracted by LDA modules facilitate both distortion-type classification and IQA of authentic images.

*Effectiveness of DTI module:* Compared with EVT+LDA+AQP and EVT+LDA+AQP+DTI (i.e., the proposed method), we can see that the DTI module makes SROCC and PLCC further improved to the highest value of 0.977, 0.978 on the LIVE dataset and 0.982, 0.982 on the CSIQ dataset, respectively. This demonstrates that the classification of distorted types by the DTI module can contribute to improvement of IQA accuracy.

*Effectiveness of AQP module:* Compared with EVT+LDA+QP+DTI and EVT+LDA+AQP+DTI, we can see from the comparison that AQP has a significant improvement for IQA results, which verifies the effectiveness of the AQP module in adaptively encoding multi-scale features for quality regression of authentic images with various contents.

## 5 Discussion

In this section, we first summarize the main differences between our proposed METER method and previous studies

Fig. 4 MSFE semantic features interpretation on KonIQ dataset

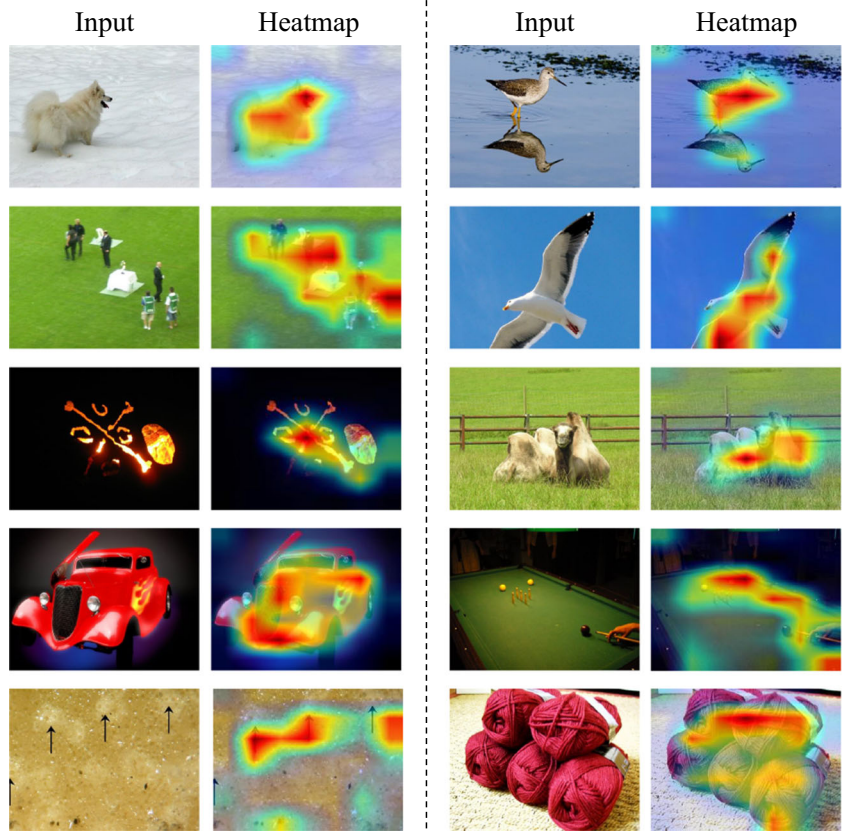


Fig. 5 AQP weights 3D space map after PCA transformation

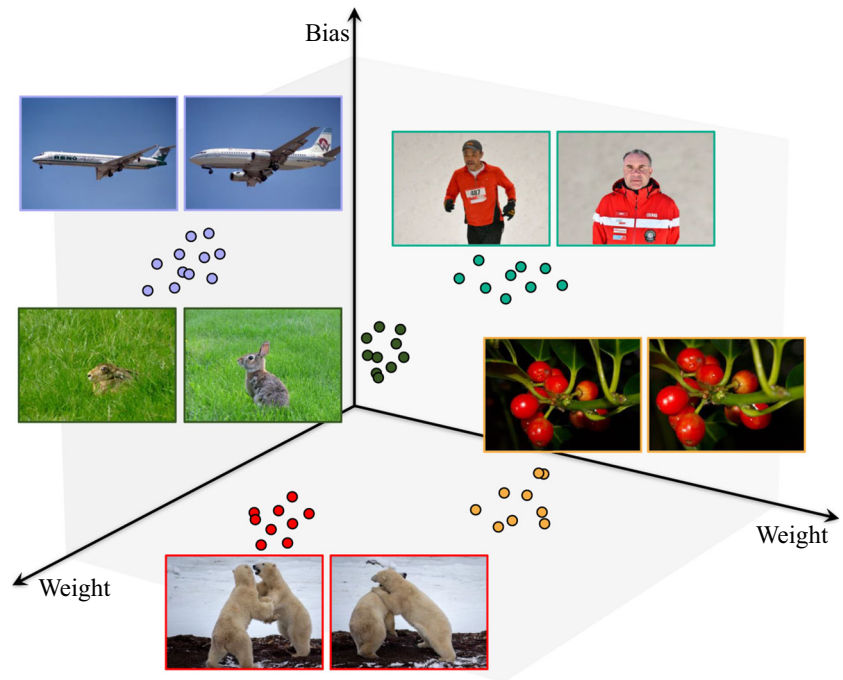




Fig. 6 Best and worst IQA results for each dataset

on NR-IQA. We also point out the limitations of our proposed method as well as potential solutions to deal with these limitations in the future.

### 5.1 Comparison with previous work

Compared with the conventional NR-IQA methods [10, 14–16, 22, 77–82] and learning-based NR-IQA methods [23–26, 30, 31, 83], METER adopts an end-to-end multi-task efficient transformer for NR-IQA, predicting the distortion type and the quality score simultaneously. Specifically, METER integrates automatic feature extraction, distortion classification and quality prediction into a unified framework, enabling these interdependent tasks to be more effectively coordinated in a task-specific manner. In addition, METER produces multi-scale attention features according to image contents and fuse distortion classification information for adaptive quality prediction, which is more consistent with the perceptual process by which the HVS evaluates image quality.

Table 7 shows the quantitative comparison results of various state-of-the-art methods, which can be divided into two

categories according to the adopted framework. METER, as a unified end-to-end optimized framework, has the advantage of incorporating both feature extraction and quality regression, making it stand out from conventional two-stage NR-IQA methods [84–87]. Unlike other NR-IQA methods that are based on a multi-task framework, such as IQA-CNN++ [37], MEON [1], DBCNN [27], GraphIQA [3], METER utilizes transformers to extract multi-scale and long-range features that cooperatively increase global and local interactions between different regions of the image, which is essential for IQA. Furthermore, METER can further improve IQA accuracy by adapting to image contents through the adjustment of weights in its fully connected layers. Although several methods, such as RankIQA [89], HyperIQA [2], and MANIQA [31], extract multi-scale features in the backbone network, none of them take the effect of image distortion types on IQA into account. Additionally, although HyperIQA [2] uses the similar adaptive fully-connected layer, it does not consider the effects of distortion types and long-range features on IQA. While transformer-based methods [30, 31, 43, 90] have been demonstrated to be effective for

Table 6 Ablation results on authentic and synthetic datasets

Components	LIVE Challenge		LIVE			CSIQ		
	SROCC	PLCC	SROCC	PLCC	ACC(%)	SROCC	PLCC	ACC(%)
EVT+DTI	–	–	–	–	97.78	–	–	97.62
EVT+LDA+DTI	–	–	–	–	<b>98.94</b>	–	–	<b>98.98</b>
EVT+AQP	0.885	0.912	0.956	0.964	–	0.962	0.969	–
EVT+LDA+AQP	<b>0.896</b>	<b>0.923</b>	0.976	0.975	–	0.972	0.976	–
EVT+LDA+QP+DTI	–	–	0.965	0.967	98.02	0.966	0.968	97.81
EVT+LDA+AQP+DTI	–	–	<b>0.977</b>	<b>0.978</b>	98.06	<b>0.982</b>	<b>0.982</b>	97.85



**Table 7** A Brief Description of the NR-IQA Studies

Reference	Framework	Methodology				Adaptation	Long-range	Datasets
		Deep features	Multi-scale	Multi-task	Adaptation			
CORNIA [84]	Feature extraction	-	-	-	-	-	LIVE, TID2008	
BRISQUE [85]	+ regression/fitting	-	✓	-	-	-	LIVE	
DeepBIQ [86]		✓	-	-	-	-	LIVEC	
DeepRN [87]		✓	-	-	-	-	KoniQ	
IQA-CNN++ [37]	End-to-end	✓	-	✓	-	-	LIVE	
WaDIQaM [69]		✓	-	-	-	-	LIVE, CSIQ, TID2013, LIVEC	
Hallucinated-IQA [88]		✓	✓	-	✓	-	LIVE, CSIQ, TID2008, TID2013	
Rankiqqa [89]		✓	-	-	-	-	TID2013, LIVE	
MEON [1]		✓	-	✓	-	-	TID2013, CSIQ	
DBCNN [27]		✓	-	✓	-	-	LIVE, CSIQ, TID2013, LIVEC	
HyperIQA [2]		✓	✓	-	✓	-	LIVE, CSIQ, LIVEC, BID, KoniQ	
Graphiqqa [3]		✓	-	✓	-	-	LIVE, CSIQ, LIVEC, Kadid, KoniQ	
TRIQ [43]		✓	-	-	-	✓	LIVE, KoniQ	
TReS [30]		✓	-	-	-	✓	LIVE, CSIQ, TID2013, KADID, KoniQ	
Maniqqa [31]		✓	✓	-	-	✓	LIVE, CSIQ, TID2013, KADID, PIPAL	
METER		✓	✓	✓	✓	✓	LIVE, CSIQ, LIVEC, BID, KoniQ	

learning long-range dependencies, METER's methodology is more comprehensive and relevant to the perceptual process of HVS, as evidenced by the comparison presented in Table 7.

## 5.2 Limitations and future work

Although METER achieves remarkable performance in authentic NR-IQA, its performance and generalizability could be further improved in the future by carefully dealing with the following limitations or challenges.

Currently, METER is designed based on transformers, which lack some of the inductive biases inherent to CNNs, such as translation equivariance and locality, thus yielding low generalizability with insufficient training data. To make full use of training data, some strong data augmentation and regularization techniques can be incorporated into METER to improve its generalizability. Furthermore, currently the identification of distortion type is used for synthetic images, ignoring the fact that the distortion type annotations of authentic images are always unavailable. To facilitate the quality regression of authentic images, an automatic and anti-label-noise labeling model can be incorporated into our framework to identify the distortion type as additional input to AQP module.

## 6 Conclusion

In this paper, we introduce an end-to-end multi-task efficient transformer (METER) to handle two challenging tasks of distortion type identification and NR-IQA. Our method METER employs a semantic feature extraction module to capture global and local features at multiple scales, incorporates distortion type identification into IQA via a multi-task learning framework, and uses an adaptive fully connected layer to construct an adaptive quality prediction module with distortion-type probabilities for accurate IQA of authentic images. Furthermore, METER can cope with IQA of authentic images with various contents and multiple distortion types effectively, affording accurate distortion type identification, and adaptive attention to various objects. Extensive experiments on benchmark datasets have shown that METER has remarkable performance and great generalizability, achieves higher efficiency than existing transformer-based methods, which is valuable for real-world scenario applications.

**Acknowledgements** This work was supported by the National Natural Science Foundation (NSF) of China under Grants 51979021 and 51709028, Natural Science Foundation of Liaoning under Grant 2019JH8/10100045, China Scholarship Council (CSC) under Grant 202206570013, Dalian High-level Talent Innovation Support Program Project 2019RQ008 and Fundamental Research Funds for the Central Universities under Grant 3132022218 and 3132019317.

**Author Contributions** Pengli Zhu designed the framework and network architecture, carried out the implementation, performed the experiments and analysed the data. Pengli Zhu and Siyuan Liu wrote the manuscript. Siyuan Liu, Yancheng Liu and Pew-Thian Yap revised the manuscript. Siyuan Liu conceived the study and were in charge of overall direction and planning.

**Availability of data and materials** the datasets generated during and/or analysed during the current study can be publicly available with links from the corresponding references.

**Code Availability** <https://github.com/Idea89560041/METER>.

## Declarations

**Conflict of interest/Competing interests** the authors declare no competing interests.

**Ethics approval** the manuscript is submitted through all the authors' consent and promises not to submit to other journals.

## References

1. Ma K, Liu W, Zhang K, Duanmu Z, Wang Z, Zuo W (2017) End-to-end blind image quality assessment using deep neural networks. *IEEE Trans Image Process* 27(3):1202–1213
2. Su S, Yan Q, Zhu Y, Zhang C, Ge X, Sun J, 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*, pp. 3667–3676
3. Sun S, Yu T, Xu J, Lin J, Zhou W, Chen Z (2022) Graphiqa: Learning distortion graph representations for blind image quality assessment. *IEEE Transactions on Multimedia*
4. Di Claudio ED, Jacovitti G (2017) A detail-based method for linear full reference image quality prediction. *IEEE Trans Image Process* 27(1):179–193
5. Sun W, Liao Q, Xue J-H, Zhou F (2018) Spsim: A superpixel-based similarity index for full-reference image quality assessment. *IEEE Trans Image Process* 27(9):4232–4244
6. Bae S-H, Kim M (2016) A novel image quality assessment with globally and locally consistent visual quality perception. *IEEE Trans Image Process* 25(5):2392–2406
7. Bampis CG, Gupta P, Soundararajan R, Bovik AC (2017) Speed-qa: Spatial efficient entropic differencing for image and video quality. *IEEE Signal Process Lett* 24(9):1333–1337
8. Min X, Gu K, Zhai G, Hu M, Yang X (2018) Saliency-induced reduced-reference quality index for natural scene and screen content images. *Signal Process* 145:127–136
9. Zhu W, Zhai G, Min X, Hu M, Liu J, Guo G, Yang X (2019) Multi-channel decomposition in tandem with free-energy principle for reduced-reference image quality assessment. *IEEE Trans Multimed* 21(9):2334–2346
10. Zhai G, Min X, Liu N (2019) Free-energy principle inspired visual quality assessment: An overview. *Digit Signal Process* 91:11–20
11. Lu Y, Li W, Ning X, Dong X, Zhang Y, Sun L (2020) Image quality assessment based on dual domains fusion. In: *2020 International Conference on High Performance Big Data and Intelligent Systems (HPBD&IS)*, pp 1–6. IEEE
12. Lu Y, Li W, Ning X, Dong X, Zhang L, Sun L, Cheng C (2021) Blind image quality assessment based on the multiscale and dual-domains features fusion. *Practice and Experience, Concurrency and Computation*, p 6177

13. Min X, Zhai G, Gu K, Fang Y, Yang X, Wu X, Zhou J, Liu X (2016) Blind quality assessment of compressed images via pseudo structural similarity. In: 2016 IEEE International Conference on Multimedia and Expo (ICME), pp 1–6. IEEE
14. Zhan Y, Zhang R (2017) No-reference jpeg image quality assessment based on blockiness and luminance change. *IEEE Signal Process Lett* 24(6):760–764
15. Dong L, Zhou J, Tang YY (2018) Effective and fast estimation for image sensor noise via constrained weighted least squares. *IEEE Trans Image Process* 27(6):2715–2730
16. Li L, Xia W, Lin W, Fang Y, Wang S (2016) No-reference and robust image sharpness evaluation based on multiscale spatial and spectral features. *IEEE Trans Multimed* 19(5):1030–1040
17. Dendi SVR, Channappayya SS (2020) No-reference video quality assessment using natural spatiotemporal scene statistics. *IEEE Trans Image Process* 29:5612–5624
18. Liu Y, Gu K, Zhang Y, Li X, Zhai G, Zhao D, Gao W (2019) Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception. *IEEE Trans Circ Syst Vid Technol* 30(4):929–943
19. Yan B, Bare B, Tan W (2019) Naturalness-aware deep no-reference image quality assessment. *IEEE Trans Multimed* 21(10):2603–2615
20. Liu Y, Yin X, Wang Y, Yin Z, Zheng Z (2022) Hvs-based perception-driven no-reference omnidirectional image quality assessment. *IEEE Trans Instrum Meas* 72:1–11
21. Yao J, Shen J, Yao C (2023) Image quality assessment based on the perceived structural similarity index of an image. *Mathematical Biosciences and Engineering: MBE* 20(5):9385–9409
22. Zhang F, Roysam B (2016) Blind quality metric for multidistortion images based on cartoon and texture decomposition. *IEEE Signal Process Lett* 23(9):1265–1269
23. Kim J, Nguyen A-D, Lee S (2018) Deep cnn-based blind image quality predictor. *IEEE Trans Neural Netw Learn Syst* 30(1):11–24
24. Wu Q, Li H, Ngan KN, Ma K (2017) Blind image quality assessment using local consistency aware retriever and uncertainty aware evaluator. *IEEE Trans Circ Syst Vid Technol* 28(9):2078–2089
25. Pang Y, Zhou B, Nie F (2019) Simultaneously learning neighborhood and projection matrix for supervised dimensionality reduction. *IEEE Trans Neural Netw Learn Syst* 30(9):2779–2793
26. Liu S, Thung K-H, Lin W, Yap P-T, Shen D (2020) Real-time quality assessment of pediatric mri via semi-supervised deep nonlocal residual neural networks. *IEEE Trans Image Process* 29:7697–7706
27. Zhang W, Ma K, Yan J, Deng D, Wang Z (2018) Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Trans Circ Syst Vid Technol* 30(1):36–47
28. Li D, Jiang T, Lin W, Jiang M (2018) Which has better visual quality: The clear blue sky or a blurry animal? *IEEE Trans Multimed* 21(5):1221–1234
29. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: *Advances in neural information processing systems*, pp 5998–6008
30. Golestaneh SA, Dadsetan S, Kitani KM (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*, pp 1220–1230
31. Yang S, Wu T, Shi S, Lao S, Gong Y, Cao M, Wang J, Yang Y (2022) Maniqa: Multi-dimension attention network for no-reference image quality assessment. In: *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp 1191–1200
32. Zhang Q, Yang Y-B (2021) Rest: An efficient transformer for visual recognition. *Adv neural inf process syst* 34:15475–15485
33. Zhang L, Zhang L, Bovik AC (2015) A feature-enriched completely blind image quality evaluator. *IEEE Trans Image Process* 24(8):2579–2591
34. Xu J, Ye P, Li Q, Du H, Liu Y, Doermann D (2016) Blind image quality assessment based on high order statistics aggregation. *IEEE Trans Image Process* 25(9):4444–4457
35. Kim J, Lee S (2016) Fully deep blind image quality predictor. *IEEE J Sel Top Signal Process* 11(1):206–220
36. Zhang Y, Chandler DM (2018) Opinion-unaware blind quality assessment of multiply and singly distorted images via distortion parameter estimation. *IEEE Trans Image Process* 27(11):5433–5448
37. Kang L, Ye P, Li Y, Doermann D (2015) Simultaneous estimation of image quality and distortion via multi-task convolutional neural networks. In: 2015 IEEE International Conference on Image Processing (ICIP), pp 2791–2795. IEEE
38. Zeng H, Zhang L, Bovik AC (2018) Blind image quality assessment with a probabilistic quality representation. In: 2018 IEEE International Conference on Image Processing (ICIP) p
39. Bahdanau D, Cho KH, Bengio Y (2015) Neural machine translation by jointly learning to align and translate. In: 3rd International Conference on Learning Representations, ICLR 2015
40. Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, Dehghani M, Minderer M, Heigold G, Gelly S, Uszkoreit J, Houshy N (2021) An image is worth 16x16 words: Transformers for image recognition at scale. *ICLR*
41. Carion N, Massa F, Synnaeve G, Usunier N, Kirillov A, Zagoruyko S (2020) End-to-end object detection with transformers. In: *European conference on computer vision*, pp 213–229. Springer
42. Chen H, Wang Y, Guo T, Xu C, Deng Y, Liu Z, Ma S, Xu C, Xu C, Gao W (2021) Pre-trained image processing transformer. In: *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp 12299–12310
43. You J, Korhonen J (2021) Transformer for image quality assessment. In: 2021 IEEE International Conference on Image Processing (ICIP), pp 1389–1393. IEEE
44. Liu J, Li X, Peng Y, Yu T, Chen Z (2022) Swiniqa: Learned swin distance for compressed image quality assessment. In: *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pp 1795–1799
45. Xu Y, Wei H, Lin M, Deng Y, Sheng K, Zhang M, Tang F, Dong W, Huang F, Xu C (2022) Transformers in computational visual media: A survey. *Comput Vis Med* 8:33–62
46. Liu Y, Zhang Y, Wang Y, Hou F, Yuan J, Tian J, Zhang Y, Shi Z, Fan J, He Z (2023) A survey of visual transformers. *IEEE Transactions on Neural Networks and Learning Systems*
47. Li K, Wang Y, Zhang J, Gao P, Song G, Liu Y, Li H, Qiao Y (2023) Uniformer: Unifying convolution and self-attention for visual recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*
48. Fan X, Liu H (2023) Flexformer: Flexible transformer for efficient visual recognition. *Pattern Recognit Lett* 169:95–101
49. Li W, Li J, Gao G, Deng W, Zhou J, Yang J, Qi G-J (2023) Cross-receptive focused inference network for lightweight image super-resolution. *IEEE Transactions on Multimedia*
50. Feng H, Wang L, Li Y, Du A (2022) Lkasr: Large kernel attention for lightweight image super-resolution. *Knowl-Based Syst* 252:109376
51. Lin X, Yu L, Cheng K-T, Yan Z (2023) Batformer: Towards boundary-aware lightweight transformer for efficient medical image segmentation. *IEEE Journal of Biomedical and Health Informatics*
52. Yang J, Tu J, Zhang X, Yu S, Zheng X (2023) Tse deeplab: An efficient visual transformer for medical image segmentation. *Biomed Signal Process Control* 80:104376
53. Zhao Z, Hao K, Liu X, Zheng T, Xu J, Cui S, He C, Zhou J, Zhao G (2023) Mcanet: Hierarchical cross-fusion lightweight transformer based on multi-convhead attention for object detection. *Image and Vision Computing*, p 104715

54. Ye T, Qin W, Zhao Z, Gao X, Deng X, Ouyang Y (2023) Real-time object detection network in uav-vision based on cnn and transformer. *IEEE Trans Instrum Meas* 72:1–13
55. Ioffe S, Szegedy C (2015) Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: *International conference on machine learning*, pp 448–456. PMLR
56. Li Y, Yuan Y (2017) Convergence analysis of two-layer neural networks with relu activation. *Advances in neural information processing systems*, vol 30
57. Han J, Moraga C (1995) The influence of the sigmoid function parameters on the speed of backpropagation learning. In: *International workshop on artificial neural networks*, pp 195–201. Springer
58. Kabani A, El-Sakka MR (2016) Object detection and localization using deep convolutional networks with softmax activation and multi-class log loss. In: *Image analysis and recognition: 13th International conference, ICIAR 2016, in Memory of Mohamed Kamel, Póvoa de Varzim, Portugal, July 13-15, 2016, Proceedings 13*, pp 358–366. Springer
59. Ulyanov D, Vedaldi A, Lempitsky V (2016) Instance normalization: The missing ingredient for fast stylization. [arXiv:1607.08022](https://arxiv.org/abs/1607.08022)
60. Xu J, Sun X, Zhang Z, Zhao G, Lin J (2019) Understanding and improving layer normalization. *Advances in Neural Information Processing Systems*, vol 32
61. Srivastava N, Hinton G, Krizhevsky A, Sutskever I, Salakhutdinov R (2014) Dropout: a simple way to prevent neural networks from overfitting. *J Mach Learn Res* 15(1):1929–1958
62. Moorthy AK, Bovik AC (2010) A two-step framework for constructing blind image quality indices. *IEEE Signal Process Lett* 17(5):513–516
63. Ghadiyaram D, Bovik AC (2015) Massive online crowdsourced study of subjective and objective picture quality. *IEEE Trans Image Process* 25(1):372–387
64. Hosu V, Lin H, Sziranyi T, Saupe D (2020) Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Trans Image Process* 29:4041–4056
65. Ciancio A, Silva EA, Said A, Samadani R, Obrador P et al (2010) No-reference blur assessment of digital pictures based on multi-feature classifiers. *IEEE Trans Image Process* 20(1):64–75
66. Thomee B, Shamma DA, Friedland G, Elizalde B, Ni K, Poland D, Borth D, Li L-J (2016) Yfcc100m: The new data in multimedia research. *Commun ACM* 59(2):64–73
67. Sheikh HR, Sabir MF, Bovik AC (2006) A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Trans Image Process* 15(11):3440–3451
68. Larson EC, Chandler DM (2010) Most apparent distortion: full-reference image quality assessment and the role of strategy. *J Electron Imaging* 19(1):011006
69. Bosse S, Maniry D, Müller K-R, Wiegand T, Samek W (2017) Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Trans Image Process* 27(1):206–219
70. Group VQE, et al (2000) Final report from the video quality experts group on the validation of objective models of video quality assessment. In: *VQEG Meeting, Ottawa, Canada, March, 2000*
71. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al (2019) Pytorch: An imperative style, high-performance deep learning library. *Adv Neural Infor Process Syst* 32:8026–8037
72. Zhang Z (2018) Improved adam optimizer for deep neural networks. In: *2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS)*, pp 1–2. IEEE
73. Deng J, Dong W, Socher R, Li L-J, Li K, Fei-Fei L (2009) Imagenet: A large-scale hierarchical image database. In: *2009 IEEE Conference on computer vision and pattern recognition*, pp 248–255. IEEE
74. Glorot X, Bengio Y (2010) Understanding the difficulty of training deep feedforward neural networks. In: *Proceedings of the 13th International conference on artificial intelligence and statistics*, pp 249–256. JMLR Workshop and conference proceedings
75. Zhang Q, Rao L, Yang Y (2021) Group-cam: Group score-weighted visual explanations for deep convolutional networks. [arXiv:2103.13859](https://arxiv.org/abs/2103.13859)
76. Tipping ME, Bishop CM (1999) Probabilistic principal component analysis. *J Royal Stat Soc Ser B (Stat Methodol)* 61(3):611–622
77. Zhang J, Le TM (2010) A new no-reference quality metric for jpeg2000 images. *IEEE Trans Cons Electron* 56(2):743–750
78. Liang L, Wang S, Chen J, Ma S, Zhao D, Gao W (2010) No-reference perceptual image quality metric using gradient profiles for jpeg2000. *Signal Process Image Commun* 25(7):502–516
79. Wang Q, Chu J, Xu L, Chen Q (2016) A new blind image quality framework based on natural color statistic. *Neurocomput* 173:1798–1810
80. Lee D, Plataniotis KN (2016) Toward a no-reference image quality assessment using statistics of perceptual color descriptors. *IEEE Trans Image Process* 25(8):3875–3889
81. Liu T-J, Liu K-H (2017) No-reference image quality assessment by wide-perceptual-domain scorer ensemble method. *IEEE Trans Image Process* 27(3):1138–1151
82. Freitas PG, Akamine WY, Farias MC (2018) No-reference image quality assessment using orthogonal color planes patterns. *IEEE Trans Multimed* 20(12):3353–3360
83. Ma K, Liu W, Liu T, Wang Z, Tao D (2017) dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs. *IEEE Trans Image Process* 26(8):3951–3964
84. Ye P, Kumar J, Kang L, Doermann D (2012) Unsupervised feature learning framework for no-reference image quality assessment. In: *2012 IEEE Conference on computer vision and pattern recognition*, pp 1098–1105. IEEE
85. Mittal A, Moorthy AK, Bovik AC (2012) No-reference image quality assessment in the spatial domain. *IEEE Trans Image Process* 21(12):4695–4708
86. Bianco S, Celona L, Napoletano P, Schettini R (2018) On the use of deep learning for blind image quality assessment. *Signal, Image Vid Process* 12(2):355–362
87. Varga D, Saupe D, Szirányi T (2018) Deepnrn: A content preserving deep architecture for blind image quality assessment. In: *2018 IEEE International Conference on Multimedia and Expo (ICME)*, pp 1–6. IEEE
88. Lin K-Y, Wang G (2018) Hallucinated-iqa: No-reference image quality assessment via adversarial learning. In: *Proceedings of the IEEE Conference on computer vision and pattern recognition*, pp 732–741
89. Liu X, Van De Weijer J, Bagdanov AD (2017) Rankiqa: Learning from rankings for no-reference image quality assessment. In: *Proceedings of the IEEE International conference on computer vision*, pp 1040–1049
90. Chen D, Wang Y, Gao W (2020) No-reference image quality assessment: An attention driven approach. *IEEE Trans Image Process* 29:6496–6506

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.





**Pengli Zhu** received the B.Eng. degree in marine engineering and the M.Eng. degree in naval architecture and ocean engineering, in 2017 and 2020, respectively, from Dalian Maritime University, Dalian, China, where he is currently working toward the Ph.D. degree in marine engineering. He is also a joint-training Ph.D. student with the College of Design and Engineering, National University of Singapore, Singapore. His current research interests include image analysis, intelligent

control, and machine learning.



**Siyuan Liu** received the B.E. degree in electrical engineering and automation, the M. E. degree in naval architecture and ocean engineering, and the Ph.D. degree in marine engineering from Dalian Maritime University, Dalian, China, in 2012, 2014, and 2018, respectively. From 2018 to 2021, he was a Post-Doctoral Research Associate with the Department of Radiology and Biomedical Research Imaging Center (BRIC), the University of North Carolina at Chapel Hill,

Chapel Hill, USA. He is currently a Professor with the Department of Marine Engineering, Dalian Maritime University, Dalian, China. His current research interests include intelligent image processing, autonomous control, and machine learning.



**Yancheng Liu** received the B.S. and M.S. degrees in electrical engineering from the Harbin Institute of Technology, Harbin, China, in 1985 and 1988, respectively, and the Ph.D. degree in marine engineering from Dalian Maritime University, Dalian, China, in 2002. He is currently a Professor with the College of Marine Engineering, Dalian Maritime University. His current research interests include image analysis, intelligent control, and machine learning.



**Pew-Thian Yap** is a Professor of the Department of Radiology and the Director of the Image Analysis Core of the Biomedical Research Imaging Center (BRIC) of the University of North Carolina at Chapel Hill. He has published more than 250 peer-reviewed articles in major international journals and conference proceedings. His research interests include medical image analysis, machine learning, and neuroscience.