This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3531416

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2024.0429000

Image quality assessment based on multi-scale representation and shifting transformer

GENG FU[1], ZIYU WANG[2], CUIJUAN ZHANG[2], ZERONG QI[3], MINGZHENG HU[3], SHUJUN FU[4], AND YUNFENG ZHANG[1]

[1]School of Computing and Artificial Intelligence, Shandong Univrsity of Finance and Economics, Jinan 250014, China
[2]Department of Interventional Therapy, Yidu Central Hospital of Weifang, Qingzhou 262500, China
[3]Shandong Chengshi Electronic Technology Limited Company, Jinan 250002, China
[4]School of Matahematics, Shandong University, Jinan 250100, China

Corresponding author: Yunfeng Zhang (e-mail: yfzhang@sdufe.edu.cn).

The research has been supported in part by the National Natural Science Foundation of China (12071263, 11971269), and the innovation ability improvement project of science and technology-based SMEs in Shandong Province (NO. 2022TSGC2072).

ABSTRACT In automatic control systems, sensors and cameras are often used to capture images of the environment or processes being monitored. The quality of these images is paramount as it directly affects the system's ability to accurately interpret and respond to the visual information. Image Quality Assessment (IQA) is a crucial metric for intelligent control systems and computer vision tasks, such as surveillance, restoration, and fingerprint identification, significantly advancing algorithm development in these areas. Recently, transformer-based algorithms have excelled in computer vision, particularly in image classification, surpassing convolutional neural network (CNN) methods. To enhance IQA using transformers, we propose Swin-MIQT, a multi-scale spatial pooling transformer with shifted windows. As a no-reference (NR) IQA method, Swin-MIQT processes images at their original resolution without resizing or cropping, unlike standard vision transformers. By using shifted windows, we reduce computational load through efficient self-attention processing. Additionally, a spatial pyramid pooling layer captures diverse image quality information, improving IQA accuracy for distorted images. Comprehensive experiments show that Swin-MIQT achieves state-of-the-art performance on three synthetic distortion databases (LIVE, LIVE MD, TID2013) and competitive results on three authentic distortion databases (LIVE Challenge, KonIQ-10K, SPAQ).

INDEX TERMS Image quality assessment, multi-scale, no-reference/blind, spatial pooling, shifted window, transformer.

I. INTRODUCTION

In the digital network monitoring and intelligent control systems, a vast number of digital images are generated daily across various electronic devices, such as smartphones, cameras, and computers [1]. However, these images are often subjected to a range of distortions during acquisition, processing, transmission, storage, and display [2]. As a result, assessing the perceptual quality of digital images becomes a crucial task. Image quality assessment (IQA) seeks to assign a quantifiable quality score to each distorted image, with this score closely correlating to human perception of image quality. IQA is widely used in many computer vision tasks: (a) quality screening of image capturing systems [3], e.g., face recognition and fingerprint recognition; (b) imaging systems [4], e.g., balancing between the used CT doses and CT image quality in low-dose CT imaging, and multi-modal image registration and fusion systems; (c) search engines regard the weighted sum of image quality and content indexes as a ranking indicator to sort the searched images, and then the image with the highest quality and most relevant content will be listed first for end-users [5]. These application scenarios enumerated above drive IQA forward and vice versa.

IQA is usually divided into subjective IQA and objective IQA. Although subjective IQA through subjective experiments can obtain the quantified quality scores of damaged images which are consistent with human visual system (HVS), it is very time-consuming and money-consuming to implement. Due to the reasons mentioned above, subjective IQA can only be applied in laboratory environment and is difficult to be used for large-scale images. The purpose of objective IQA is to design a computational model that is consistent with HVS, where the model can automatically evaluate the quality

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3531416

of images. Accordingly, considering economy and efficiency, objective IQA is of great research value. In recent years, although many objective IQA models [6]–[14] have been proposed to promote the development of IQA community, and can achieve good results for dealing with natural images on synthetic distortions, they perform poorly for authentically distorted and content-specific images. According to the used information amount of each reference image, these models can be divided into three types: full-reference IQA (FR-IQA) for the whole information of each reference image used, reduced-reference IQA (RR-IQA) for part information of each reference image used, and no-reference/blind IQA (NR-IQA/BIQA) for without reference images used. In fact, most of the images we receive and send have no the corresponding pristine reference images, which makes NR-IQA more practical and valuable. Nowadays, NR-IQA is a research focus in the field of IQA. Notably, for quantifying the quality of some content-complex and scene-specific images which have the corresponding reference images, FR-IQA can perform better than NR-IQA.

In the era of machine learning, the blood of IQA task is the labeled databases with mean opinion scores/differential mean opinion scores (MOSs/DMOSs), and the evolution of these databases reflects the development trend and research status of IQA industry.

The early IQA databases are artificially synthesized, and their sizes are small, e.g., IVC [15], LIVE [16], and CSIQ [17]. Models designed and trained on these synthetic images are limited and their test results on real-world distorted images are poor. In order to improve the performance of models tackling authentic images, some large-scale databases in the wild have been created recently, e.g., KonIQ-10K [18], SPAQ [19], and PaQ-2-PiQ [20], which bring a great challenge to IQA. Some NR-IQA approaches use hand-crafted features [21], [22] or learned features from convolutional neural networks (CNNs) [23], [24] to represent the perceptual quality for distorted images, and then project these extracted features into quality scores by utilizing support vector regression (SVR). Some NR-IQA approaches avoid using SVR to achieve end-to-end architecture, e.g., CNN-based models [1], [25], [26], ranking-based models [27]–[29] for mitigating the shortage of available training data, and generative adversarial networks (GANs) based models [30], [31].

Recently, transformer-based algorithms [32], [33] have attained competitive performance on IQA. They have a common merit that multi-resolution images can be received as inputs without cropping or resizing. Transformer [34] is characterized by self-attention mechanism which computes the interaction of image patches. How to improve the predicted accuracy further in accordance with HVS highly is very important.

In this paper, we introduce a multi-scale spatial pooling image quality transformer, named Swin-MIQT, designed to quantify the quality of distorted images in an end-to-end manner using shifted windows. Swin-MIQT demonstrates impressive results across six public IQA databases. For image inputs, we initially select ResNet50 [35] as the image encoding module to obtain embedded features. Subsequently, we employ a spatial pyramid pooling layer [36] to generate multi-scale representations of these embedded features. To mitigate the computational complexity of the transformer, we conduct multi-head self-attention (MHSA) processing within each shifted window, enabling the model to learn global features related to perceptual quality [37]. For six benchmark databases, our proposed model Swin-MIQT achieves both state-of-the-art (SOTA) performance on three synthetically distorted databases (LIVE [16], LIVE MD [38], and TID2013 [39]) and competitive performance on three authentically distorted databases (LIVE Challenge [40], KonIQ-10K [18], and SPAQ [19]).

The remainder of this article is organized as follows. Section 2 reviews the latest research findings on IQA, especially transformer-based variants for IQA. Section 3 details the architecture of proposed model Swin-MIQT. Section 4 details the experimental results, and conducts ablation experiments to show the impacts of loss functions, multi-scale representations, and hyper-parameters. Finally, we conclude the whole article in Section 5.

II. RELATED WORKS

With the rapid development of semi-conductor industry, the cost of computing power is keeping down, which makes big models popular to deal with practical problems. In this section, we detail the latest research findings on perceptual IQA.

Recently, some models based on weight distribution [26], [41], [42] are proposed to tackle both synthetically and authentically distorted images. Most of them utilize a deep convolution network to learn perceptual quality representation for images, and then the learned representation is projected into a quality score by using a multi-layer perceptron (MLP) module. However, the convolution operation can only capture fine-grained feature and cannot capture coarse-grained feature for images. In order to solve the limitation of convolutional operation, self-attention mechanism derived from transformer [34] is introduced to extract both local and global features. Transformer [34] is a natural language processing (NLP) model proposed in 2017, and it has now become a paradigm for language modeling, image classification, object detection, and semantic segmentation. Since transformer was introduced into vision tasks, many representative models have been produced, such as DETR [43], ViT [44], DeiT [45], and Swin Transformer [37]. Transformer-based models have shown strong performance, and broken the situation of CNNs dominating visual tasks.

IQT [46] is a champion model of the NTIRE 2021 Challenge on perceptual IQA at CVPR 2021 [47]. The challenge uses expanded PIPAL database [48] as benchmark, which contains outputs of GAN-based image restoration or GAN-based compression algorithms. IQT [46] uses Inception-Resnet-V2 [49] pre-trained on ImageNet [50] with fixed weights to extract image features, and then uses classical en-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3531416

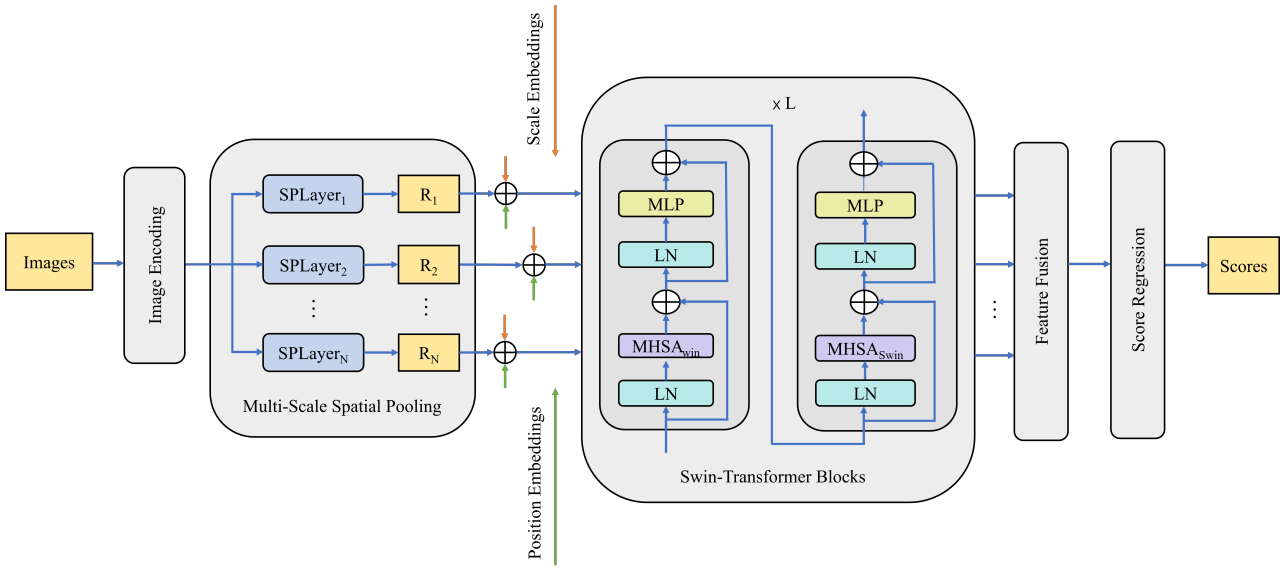

FIGURE 1: Overview of proposed Swin-MIQT. SPLayer and R mean spatial pooling layer and representation modules respectively.

TABLE 1: Details regarding six databases utilized.

Database	Year	Distorted type	No. of reference images	No. of distorted images	No. of distortion types	Distorted types of each image	Degraded levels of each distortion type	Rating distribution	MOS/DMOS
LIVE [16]	2006	Synthetic	29	779	5	1	5~7	No	DMOS[0,100]
LIVE MD [38]	2012	Synthetic	15	480	3	2	4	Yes	DMOS[0,100]
TID2013 [39]	2013	Synthetic	25	3,000	24	1	5	No	MOS[0,9]
LIVE Challenge [40]	2016	Authentic	0	1,169	N/A	N/A	N/A	No	MOS[0,100]
KonIQ-10K [18]	2018	Authentic	0	10,073	N/A	N/A	N/A	Yes	MOS[1,5]
SPAQ [19]	2020	Authentic	0	11,125	N/A	N/A	N/A	No	MOS[0,100]

coder, decoder, and MLP head modules to predict image quality. TRIQ [32] uses ResNet50 as the convolutional backbone, and achieves arbitrary resolution inputs by adopting adaptive position encoding based on transformer. MUSIQ [33] predicts quality score for each image by using the corresponding multi-scale inputs, which include the original image and its variants of different resolutions with the same aspect ratio. MANIQA [42] utilizes transposed attention and scale swin transformer blocks to strengthen global and local interaction of extracted features.

III. SWIN-MIQT FOR IMAGE QUALITY ASSESSMENT

In this section, we first depict the proposed NR-IQA model, Swin-MIQT, which is illustrated in Fig. 1, and then detail the loss functions which we used for training Swin-MIQT.

A. MODEL ARCHITECTURE

Swin-MIQT is a transformer-based variant which mainly consists of image encoding (IEncoding), multi-scale spatial pooling (MSPooling), swin-transformer blocks (SBlocks), feature fusion (FFusion), and score regression (SRegression) modules. These modules will be detailed below, along the pipeline of the proposed model.

For each received image $X \in R^{C\times H\times W}$, ResNet50 followed by a 1×1 convolutional layer is used as image encoding module to embed patch-wise information, where the 1×1 convolution aims to project the channel number of ResNet50's output into D. The embedded feature is denoted by

$$F_1 = f_{IEncoding}(X) = Conv_{1\times1}(ResNet_{50}(X)). \quad (1)$$

Multi-scale spatial pooling module consists of N spatial pooling layers (SPLayers) in a parallel manner, where each layer achieves one different scale representation. The i-th representation is denoted by

$$R_i = SPLayer_i(F_1). \quad (2)$$

The i-th spatial pooling layer $SPLayer_i$ projects F_1 into a resolution-specific representaion $R_i \in R^{D\times H_i\times W_i}$. In order to remain the positional information of distorted image, we follow traditional transformer adding the position embeddings for each representation. We follow MUSIQ [33] to define a learnable position matrix $M \in R^{G\times G}$, where each element in M is a vector of D dimensions. For each pixel-wise patch at position (j,k) in R_i, it will be projected into M proportionably at position (m_j, m_k), namely satisfying

$$\frac{j}{H_i} = \frac{m_j}{G}, \frac{k}{W_i} = \frac{m_k}{G}. \quad (3)$$

In order to remain scale information inspired by MUSIQ, we define N learnable scale vectors with D dimensions for N representations respectively. And then the scale embeddings

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3531416

and position embeddings are added into the corresponding representations. Each multi-scale representation is flattened along the height and width to attain an $(H_i \cdot W_i) \times D$ feature. Swin-transformer blocks module is a stack of L swin-transformer blocks. For an input $I \in R^{n\times d}$, the operation of self-attention (SA) in swin-transformer is denoted by

$$SA(Q,K,V) = Softmax(\frac{QK^\top}{\sqrt{d}} + B)V, \tag{4}$$

where $(Q,K,V) = I(W_Q, W_K, W_V)$, $W_Q, W_K, W_V \in R^{d\times d}$, and $B \in R^{n\times n}$ is a learnable bias of relative position. And the operation of MHSA can be denoted by

$$MHSA(I) = Concat(Head_1, \cdots, Head_k)W_C, \tag{5}$$

where $Head_i = SA(Q_i, K_i, V_i)$, $Concat$ means concatenating k heads along width, and W_C means a linear layer.

One swin-transformer block consists of two successive parts that one operation is based on regular window partition and the other operation is based on shifted window partition. For input X_1, the operation based on regular window partition is shown below.

$$Y_1 = X_1 + MHSA_{win}(LN(X_1)), \tag{6}$$

$$Z_1 = Y_1 + MLP(LN(Y_1)). \tag{7}$$

And the operation based on shifted window partition is shown below.

$$Y_2 = Z_1 + MHSA_{Swin}(LN(Z_1)), \tag{8}$$

$$Z_2 = Y_2 + MLP(LN(Y_2)). \tag{9}$$

Where *win* means that the MHSA operation is conducted within each regular window, and *Swin* means that the MHSA operation is conducted within each shifted window.

Feature fusion module consists of N parallel convolutions that convert the channels of SBlocks module's outputs to 8, where each convolution is followed by one adaptively average pooling to attain a feature of 4×4 size. Then all scale features are concatenated along the channel dimension. For a set of inputs $\{T_1, \cdots, T_N\}$, feature fusion module is formulated as

$$Y = Concat(T_1^1, \cdots, T_N^1), \tag{10}$$

where $T_i^1 = AvgPool_{4\times4}(Conv_i(T_i))$. Finally, in order to project the fused feature Y into a quality score, a two-layer MLP is proposed as score regression module, where hidden layer is followed by a GELU activation function [51], and the extended coefficient of hidden layer is set to 3. SRegression module can be formulated as

$$O = Linear_2(GELU(Linear_1(Y))). \tag{11}$$

B. LOSS COMPUTING

Loss functions are regarded as a metric to evaluate the distance between labels and model outputs, and they guide models to converge. Nowadays, two kinds of loss functions are utilized in IQA [18]. The first kind aims to train models by predicting a MOS/DMOS for each image, e.g., mean absolute error (MAE) and mean square error (MSE). The second kind aims to train models by predicting a distribution of ratings for each image, e.g., Huber loss (HLoss) [24], cross entropy, and Earth Mover's Distance (EMD) [52].

Considering that MAE is similar to MSE, we only use MAE as the loss function to train proposed model for predicting MOSs/DMOSs. Since Huber loss makes a significant progress of model performance in [24], we use Huber loss to train the proposed model for predicting rating distributions on KonIQ-10K in ablation study, which aims to research the impact of different loss functions. For image labels (MOSs/DMOSs) $\{x_1, \cdots, x_m\}$ and predicted scores $\{y_1, \cdots, y_m\}$, MAE is formulated as

$$MAE = \frac{1}{m}\sum_{i=1}^{m}|x_i - y_i|, \tag{12}$$

where m is the size of mini-batch. Huber loss for a scalar error is denoted by

$$h_\delta(x) = \begin{cases} \frac{x^2}{2} & , \text{ if } x \le \delta, \\ \delta\cdot(|x| - \frac{\delta}{2}), & otherwise, \end{cases} \tag{13}$$

where the degree of loss is contorlled by δ. Huber loss is effective for limiting larger error. For ground-truth distributions of ratings $\{q_1, \cdots, q_m\}$ and predicted distributions $\{p_1, \cdots, p_m\}$, the Huber loss is formulated as

$$HLoss = \frac{1}{mn}\sum_{i=1}^{m}\sum_{j=1}^{n}(h_\delta(q_i^j - p_i^j)), \tag{14}$$

where $q_i = \{q_i^1, \cdots, q_i^n\}$, $p_i = \{p_i^1, \cdots, p_i^n\}$, m represents the size of mini-batch, and n represents the dimentions of each rating distribution.

IV. EXPERIMENTS AND ANALYSIS

In this section, we first introduce six benchmark databases, implementation details, and performance metrics. Subsequently, the experimental results are contrasted with the current state-of-the-art (SOTA) models in Blind Image Quality Assessment (BIQA). Finally, we conduct ablation study to show the impacts of different loss functions and multi-scale representations, as well as choice of some hyper-parameters.

A. BENCHMARK DATABASES

In order to evaluate the performance of proposed Swin-MIQT, we run experiments on six benchmark databases which include three synthetic databases (LIVE [16], LIVE MD [38], and TID2013 [39]) and three authentic databases (LIVE Challenge [40], KonIQ-10K [18], and SPAQ [19]).

LIVE [16] consists of 29 pristine reference images and 779 singly distorted images. These distorted images include

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3531416

TABLE 2: Results of SROCC and PLCC on three synthetically benchmark databases. The top two results on each database are highlighted in boldface.

SROCC	LIVE [16]	LIVE MD [38]	TID2013 [39]
BRISQUE [22]	0.939	0.886	0.604
CORNIA [21]	0.947	0.899	0.678
M3 [53]	0.951	0.892	0.689
Kang-CNN [25]	0.956	**0.933**	0.558
IL-NIQE [6]	0.902	0.902	0.521
HOSA [7]	0.946	0.913	0.735
BIECON [8]	0.961	0.909	0.717
ResNet50 [35]	0.950	0.909	0.712
DIQaM-NR [26]	0.960	0.906	**0.835**
WaDIQaM-NR [26]	0.954	—	0.761
FRIQUEE [9]	0.940	0.923	0.680
RankIQA [29]	**0.981**	0.908	0.780
MEON [10]	0.943	—	0.808
DB-CNN [1]	0.968	0.927	0.816
HyperIQA [41]	0.962	—	—
Swin-MIQT	**0.984**	**0.952**	**0.863**
PLCC	LIVE	LIVE MD	TID2013
BRISQUE [22]	0.935	0.917	0.694
CORNIA [21]	0.950	0.921	0.768
M3 [53]	0.950	0.919	0.771
Kang-CNN [25]	0.953	0.927	0.653
IL-NIQE [6]	0.908	0.914	0.648
HOSA [7]	0.947	0.926	0.815
BIECON [8]	0.962	0.933	0.762
ResNet50 [35]	0.954	0.920	0.756
DIQaM-NR [26]	0.972	0.931	0.855
WaDIQaM-NR [26]	0.963	—	0.787
FRIQUEE [9]	0.944	**0.934**	0.753
RankIQA [29]	**0.982**	0.929	0.793
MEON [10]	0.954	—	—
DB-CNN [1]	0.971	**0.934**	**0.865**
HyperIQA [41]	0.966	—	—
Swin-MIQT	**0.986**	**0.940**	**0.881**

TABLE 3: Results of SROCC and PLCC on database LIVE Challenge. The top three results on each metric are highlighted in boldface.

Model	SROCC	PLCC
BRISQUE [22]	0.608	0.629
CORNIA [21]	0.629	0.671
M3 [53]	0.607	0.630
Kang-CNN [25]	0.516	0.536
IL-NIQE [6]	0.594	0.589
HOSA [7]	0.640	0.678
BIECON [8]	0.595	0.613
ResNet50 [35]	0.819	0.849
DIQaM-NR [26]	0.606	0.601
WaDIQaM-NR [26]	0.671	0.680
FRIQUEE [9]	0.682	0.705
RankIQA [29]	0.641	0.675
MEON [10]	0.688	0.693
DB-CNN [1]	**0.851**	**0.869**
HyperIQA [41]	**0.859**	**0.882**
MetaIQA [11]	0.802	0.835
Swin-MIQT	**0.841**	**0.869**

169 JPEG compressed (JPEG) images, 175 JPEG2000 compressed (JP2K) images, 145 Gaussian blur (GB) images, 145 white noise (WN) images, and 145 images of bit errors in JP2K bit stream, i.e., 779 distorted images $=$ 169 JPEG images $+$ 175 JP2K images $+$ 145 GB images $+$ 145 WN images $+$ 145 bit-error images. Each image in LIVE has a DMOS in the range $[0, 100]$, and lower DMOS corresponds to higher image quality. LIVE MD [38] consists of 2 groups of 480 multiply distorted images generated from 15 pristine reference images. The first group of images are generated by blurring and then JEPG compressing, and the second group of images are generated by blurring and then noising, where each type of distortion has 4 degraded levels containing 0 level for no distortion, i.e., 480 distorted images $= 15\times4\times4+15\times4\times4$. Each image in LIVE MD has a DMOS in the range $[0, 100]$. TID2013 [39] consists of 25 pristine reference image and $3,000$ singly distorted images. These distorted images are derived from 24 distorted types at 5 different degraded levels, i.e., $3,000$ distorted images $=$ 25 pristine reference images $\times$ 24 distorted types $\times$ 5 degraded levels. Each image in TID2013 has a MOS in the range $[0, 9]$, and higher MOS corresponds to higher image quality.

TABLE 4: Results of SROCC and PLCC on database KonIQ-10K. The best result on each metric is highlighted in boldface.

Model	SROCC	PLCC
DIIVINE [54]	0.589	0.612
BRISQUE [22]	0.705	0.707
CORNIA [21]	0.780	0.795
Kang-CNN [25]	0.572	0.584
IL-NIQE [6]	0.501	0.537
HOSA [7]	0.805	0.813
BIECON [8]	0.618	0.651
WaDIQaM-NR [26]	0.797	0.805
DB-CNN [1]	0.875	0.884
HyperIQA [41]	0.906	0.917
MetaIQA [11]	0.850	0.887
Swin-MIQT	**0.917**	**0.934**

LIVE Challenge [40] consists of $1,169$ authentically distorted images captured from some representative mobile devices. Each image in LIVE Challenge has a MOS in the range $[0, 100]$, and the acquisition of MOSs is based on a subjective quality assessment of over $8,100$ subjects. KonIQ-10K [18] consists $10,073$ authentically distorted images with MOSs obtained from a subjective quality assessment of $1,459$ subjects. Each image in KonIQ-10K has a 5-scale distribution of ratings. SPAQ [19] consists of $11,125$ realistically distorted images taken by 66 smartphones, and each image has a MOS in the range $[0, 100]$. The overall information of six IQA databases mentioned above is summarized in Table 1.

B. IMPLEMENTATION AND PERFORMANCE

We split each database into two non-overlapping subsets by following previous literature [1], [33]. One subset including 80% data is regarded as training set, and the other subset including 20% data is regarded as testing set. In order to train a content-independent model, we split each database according to pristine reference images for LIVE, TID2013, and LIVE MD , which aims to attain non-overlapping content subsets. For databases LIVE Challenge, KonIQ-10K, and SPAQ, we directly divide each database according to the principle of training/testing $= 4/1$. To the best of our knowl-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3531416

TABLE 5: Results of SROCC and PLCC on database SPAQ. The top two results on each metric are highlighted in boldface.

Model	SROCC	PLCC
DIIVINE [54]	0.599	0.600
BRISQUE [22]	0.809	0.817
CORNIA [21]	0.709	0.725
QAC [55]	0.092	0.497
IL-NIQE [6]	0.713	0.721
ResNet50 [35]	**0.908**	0.909
FRIQUEE [9]	0.819	0.830
DB-CNN [1]	**0.911**	**0.915**
Swin-MIQT	**0.908**	**0.913**

TABLE 6: Results of SROCC on comparing generalization ability of models. The best result in each column is highlighted in boldface.

Trained set	LIVE	TID2013
Testing set	TID2013	LIVE
BRISQUE [22]	0.358	0.790
CORNIA [21]	0.360	0.846
M3 [53]	0.344	0.873
HOSA [7]	0.361	0.846
DIQaM-NR [26]	0.392	–
WaDIQaM-NR [26]	0.462	–
FRIQUEE [9]	0.461	0.755
DB-CNN [1]	0.524	**0.891**
Swin-MIQT	**0.533**	0.823

edge, the evenly distributed datasets are critical to perform model potential. We use fixed training and testing sets to evaluate our proposed model, which aims to adjust hyper-parameters according to eliminating interference from uneven distribution of data, and keep data distribution even.

ResNet50 pretrained on ImageNet is used as image encoding module. Although transformer-based models can process arbitrary resolution images, they cannot conduct batch training directly. The common practice is padding or cutting each sequence of patches to an identical length. we randomly crop multiple resolution-specific patches from one image for the databases that contain images of different resolutions, which aims to achieve batch training. For small size images, we use two-scale spatial pooling operations, i.e., S1=6 and S2=12, where S1=6 means the resolution of first representation is 6×6, and S2=12 means the resolution of second representation is 12×12. For large size images, we use three-scale spatial pooling operations, i.e., S1=7, S2=14, and S3=21. Since self-attention processing is conducted within each window, multi-scale representations derived from the output of multi-scale spatial pooling module need to meet a condition, which these representations are divisible by window size. For databases LIVE, TID2013, and LIVE Challenge, we use two-scale spatial pooling operations with window size 6. For databases LIVE MD, KonIQ-10K, and SPAQ, we use three-scale spatial pooling operations with window size 7. For database LIVE, each image provides one random patch of 420×420 pixels in training phase and three random patches of 420×420 pixels in testing phase. Images share ground-truth lebels with the corresponding patches. The average predicted

TABLE 7: Results of SROCC and PLCC under different loss functions on database KonIQ-10K. The best result in each column is highlighted in boldface.

Model	SROCC	PLCC
Swin-MIQT (HLoss)	0.913	0.931
Swin-MIQT (MAE)	**0.917**	**0.934**

TABLE 8: SROCC and PLCC results on database SPAQ. The best result in each column is highlighted in boldface.

Scale	LIVE MD		KonIQ-10K	
	SROCC	PLCC	SROCC	PLCC
[21]	0.945	0.937	0.912	0.929
[21,14]	0.944	0.938	0.910	0.928
[21,14,7]	**0.952**	**0.940**	**0.917**	**0.934**

score of these three patches serves as the ultimate predicted quality score of corresponding image. For database SPAQ, each image has a resolution in the range from $1,080 \times 1,080$ pixels to $5,488 \times 6,656$ pixels. In order to control the amount of computation for affording model training, we resize each image to $1,080 \times 1,080$ pixels. AdamW optimizer with cosine warm-up is set for all training processes. Learning rate is set to 0.00001 for LIVE, LIVE MD, TID2013, KonIQ-10K, SPAQ, and 0.0001 for LIVE Challenge. (batch size, epochs) is set to (48,100), (20,15), (46,10), (38,100), (22,100), (14,100) for LIVE, LIVE MD, TID2013, LIVE Challenge, KonIQ-10K, and SPAQ, respectively. We use MAE as loss function for these six benchmark databases by default.

Two metrics are used to evaluate the consisitency between image labels and predicted scores commonly. Spearman's rank-ordered correlation coefficient (SROCC) is a metric that quantifies the monotonicity of predicting, and Pearson's linear correlation coefficient (PLCC) is a metric that quantifies the accuracy of predicting. For two one-dimensional arrays, $\{x_i\}_{i=1}^{n}$ and $\{y_i\}_{i=1}^{n}$, their SROCC can be denoted by

$$SROCC = 1 - \frac{6}{n\left(n^2-1\right)} \sum_{i=1}^{n} \left(r_{x_i} - r_{y_i}\right)^2, \qquad (15)$$

where r_{x_i} and r_{y_i} mean the ordered numbers of x_i and y_i in their repective arrays. And the PLCC can be denoted by

$$PLCC = \frac{1}{n-1} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right), \qquad (16)$$

where $\bar{x}$ and $\bar{y}$ are means of variables x and y respectively, and σ_x and σ_y are standard deviations of variables x and y respectively. For these two metrics, higher value corresponds to better model performance. A good IQA model should have SROCC and PLCC that are close to 1.

C. EXPERIMENTAL RESULTS

In order to show the powerful performance of our proposed model Swin-MIQT, we conduct some experiments on synthetically and authentically benchmark databases. There are two factors we have to take into account before comparing experimental results with current SOTA BIQA models.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3531416

TABLE 9: SROCC and PLCC results on LIVE MD using three-scale representations. The two best results are highlighted in boldface.

No.	L	K	D	SROCC	PLCC
1	[1,1,3,1]	[3,6,12,24]	96	0.916	0.908
2	[1,1,3,1]	[3,6,12,24]	192	0.909	0.899
3	[1,1,3,1]	[3,6,12,24]	384	0.922	0.911
4	[1,1,3,1]	[3,6,12,24]	576	0.928	0.914
5	[1,1,9,1]	[3,6,12,24]	384	**0.952**	**0.940**
6	[1,1,9,1]	[3,6,12,24]	576	**0.949**	**0.943**
7	[3,3,3,3]	[6,6,6,6]	384	0.924	0.913
8	[3,3,3,3]	[6,6,6,6]	576	0.943	0.943

Among these SOTA BIQA models, CNN-based models are difficult to reproduce the good results showed in respectively original papers and reproducing always descends model performance. Although reproducing is the best way to validate model performance, we respect and report the results that are reported and validated extensively in previous papers. Then, three ablation experiments are conducted for showing the effectiveness of multi-scale designing, and determining model hyper-parameters and default loss function.

For validating the performance of Swin-MIQT, we conduct experiments compared with eighteen current SOTA BIQA models, which include DIIVINE [54], BRISQUE [22], CORNIA [21], QAC [55], M3 [53], Kang-CNN [25], IL-NIQE [6], HOSA [7], BIECON [8], ResNet50 [35], DIQaM-NR [26], WaDIQaM-NR [26], FRIQUEE [9], RankIQA [29], MEON [10], DB-CNN [1], HyperIQA [41], and MetaIQA [11]. The numerical results of referenced models are reported by [1], [11], [18], [19], [41], [56], [57].

1) Results on synthetically distorted databases

Table 2 shows the results of SROCC and PLCC on three synthetically distorted benchmark databases. As listed in Table 2, Swin-MIQT outperforms the other BIQA models. Compared to ResNet50, Swin-MIQT averagely improves the values of SROCC and PLCC about 8% and 6%, respectively. The increases of performance show the effectiveness of additional modules after image encoding module. Notably, deep learning based models attain competitively performance, especially models RankIQA and DB-CNN. In order to further perceive the results of Swin-MIQT visually, some scatter diagrams are made in Fig. 2. In each scatter diagram, the axis of objective score represents predicted scores, and the axis of subjective score represents MOSs/DMOSs. Each scatter point represents an image, and a curve is fitted for all scatter points in diagram by using a non-linear logistic regressive function [16] denoted by

$$Quality(x) = \beta_1 \left(\frac{1}{2} - \frac{1}{1 + e^{\beta_2 (x - \beta_3)}}\right) + \beta_4 x + \beta_5, \quad (17)$$

where five parameters (β_1, β_2, β_3, β_4, and β_5) are computed according to all scatter points.

We plot a scatter diagram for each training set and testing set. As shown in Fig. 2, the curves on training sets are consistent with the curves on testing sets, and all curves are fitted well with the scatter points. These two facts demonstrate that Swin-MIQ has powerful learning and generalization abilities on synthetically distorted databases.

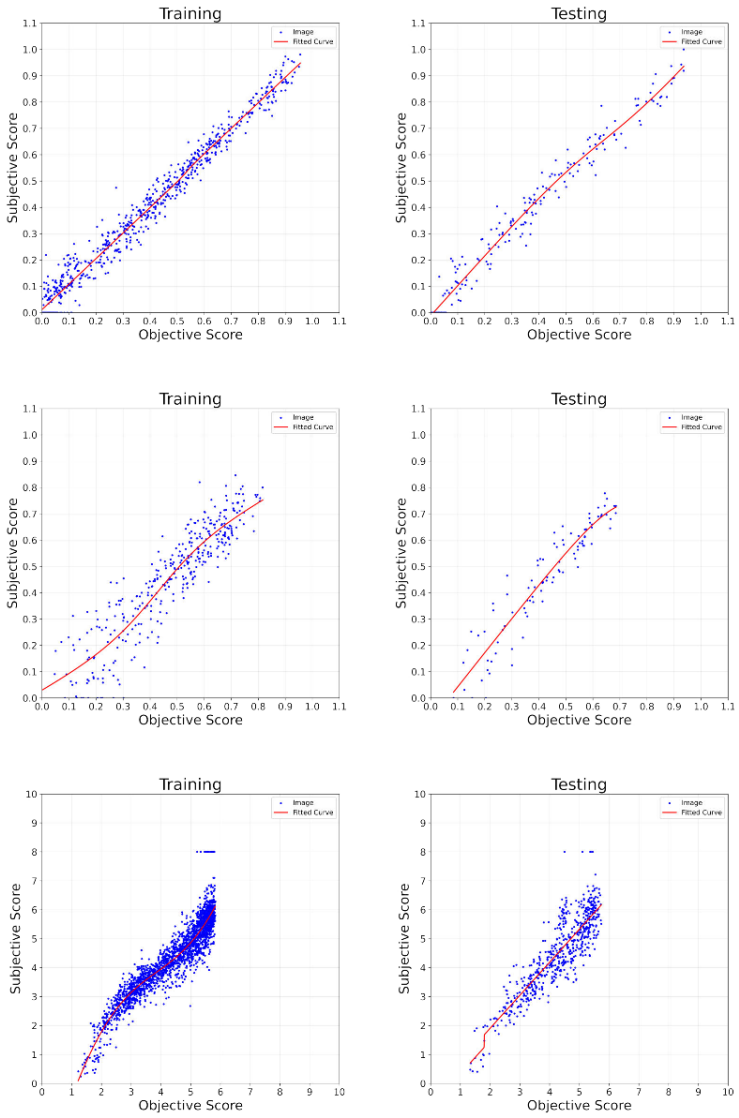

FIGURE 2: Experimental rusults of Swin-MIQT on three synthetically distorted benchmark databases (from the first to the third lines): LIVE, LIVE MD and TID 2013.

2) Results on authentically distorted databases

Table 3, 4, and 5 show the results on databases LIVE Challenge, KonIQ-10K, and SPAQ. Our proposed model achieves the top performance on these three authentically distorted databases. Specifically, on database KonIQ-10K, Swin-MIQT outperforms the other models by a large margin. On database SPAQ, Swin-MIQT has a tiny gap with the best results. Some scatter diagrams are plotted in Fig. 3 to visually perceive the numerical results of Swin-MIQT on three authentic databases.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3531416

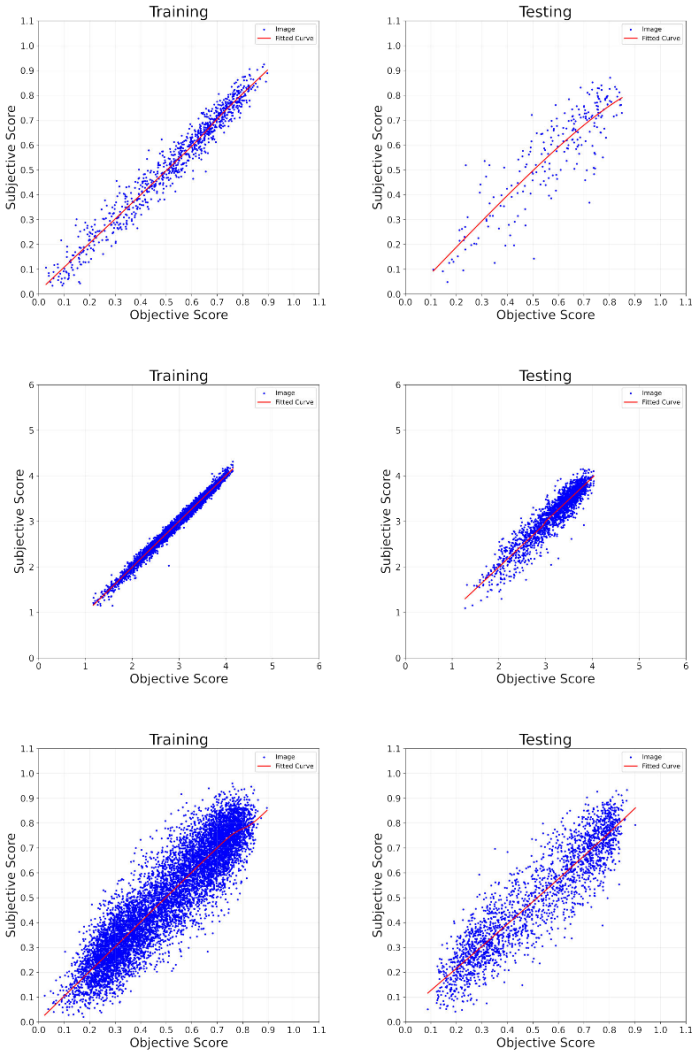

FIGURE 3: Experimental rusults of Swin-MIQT on three authentically distorted benchmark databases (from the first to the third lines): LIVE challenge, KonIQ-10K, SPAQ.

3) Comparing on generalization ability of models

An excellent model should have a good generalization ability. Nowadays, the generalization performance of learning-based models is still at low level. How to improve their generalization ability on unknown distortions is a challenge. Table 6 shows the evaluated results of cross-database, where all models are trained on LIVE or TID2013, and then tested on the other entire database. We only conduct experiments between synthetical databases, since there is a big gap on distortions between synthetical database and authentical database, which makes it difficult to perform models well [1]. As listed in Table 6, some findings can be attained. TID2013 has more distorted types than LIVE, so the results trained on TID2013 and tested on LIVE are superior to those trained on LIVE and tested on TID2013. For the result of training on LIVE and testing on TID2013, Swin-MIQT achieves the best SROCC. Although Swin-MIQT is inferior to other models on result of training on TID2013 and testing on LIVE, it is still able to attain a high SROCC. For visually perceiving the cross-database results of Swin-MIQT, two scatter diagrams are plotted in Fig. 4.

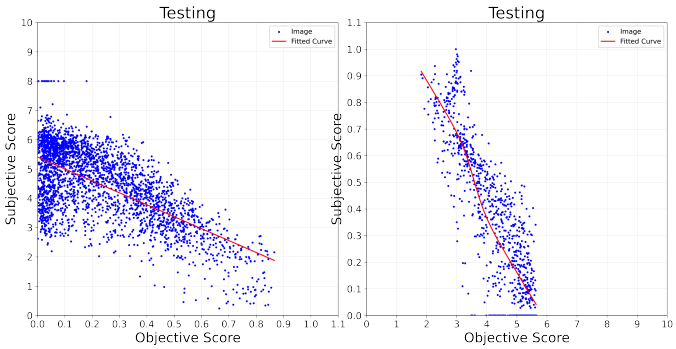

FIGURE 4: Cross-database evaluation for Swin-MIQT between LIVE and TID2013. (a) Testing results on entire TID2013 using Swin-MIQT trained on LIVE. (b) Testing results on entire LIVE using Swin-MIQT trained on TID2013.

4) Ablation study

Impacts of different loss functions. The results of SROCC and PLCC are improved a lot according to predict rating distributions, which was reported by [24]. In this part, instead of directly predicting MOSs for images, we use Huber loss function to train Swin-MIQT that aims to predict a distribution of ratings for each image in KonIQ-10K, where the parameter δ is set to $\frac{1}{9}$ as done in [24]. The output layer of Swin-MIQT uses 5 neural units followed by a Softmax operation that normalizes the 5 outputs to a five-dimensional vector of length 1 for each image, i.e., $\sum_{i=1}^{5} p_i = 1$. The objective score for each image can be formulated as

$$p = \sum_{i=1}^{5} i \cdot p_i. \tag{18}$$

MAE is the default loss function for training Swin-MIQT. The results under Huber loss and MAE are shown in Table 7. As listed in Table 7, Swin-MIQT has a better performance under MAE than Huber loss.

Impacts of different multi-scale representations. In order to validate the importance of multi-scale representations, three different compositions of scale are tested on LIVE MD and KonIQ-10k. The results of SROCC and PLCC are shown in Table 8. When Swin-MIQT is trained on LIVE MD and KonIQ-10K by using three-scale representations, it has a better performance than those using single-scale representation and two-scale representations. It means that multi-scale representations can capture more quality information at different granularities to further improve model performance.

Choice of hyper-parameters in swin-transformer blocks module. Some hyper-parameters in swin-transformer blocks need to be determined. The hyper-parameters are

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3531416

denoted by L for the number of swin-transformer blocks, K for the number of heads in swin-transformer block, and D for the channel number of scale representation. Following the philosophy of designing in [37] and [58], some combinations of L, K, and D are tested on LIVE MD, which aims to research the interaction of hyper-parameters and find an optimal combination for Swin-MIQT. Three-scale representations $[21, 14, 7]$ are applied to proposed model throughout those experiments of hyper-parameter choosing. Table 9 shows the results of SROCC and PLCC, where all blocks are split into four stages. We first increase the value of D from 96 to 576, and keep $L = [1, 1, 3, 1]$ and $K = [3, 6, 12, 24]$ fixed, which corresponds to No. 1, No. 2, No. 3, and No. 4 in Table 9. When $D = 384$ or 576, Swin-MIQT achieves a promising result, so we increase the values in L and adjust the values in K to make our proposed model perform better. Although No. 5 and No. 6 are the two best compositions, No. 5 has a smaller D, which indicates No. 5 consumes fewer resources than No. 6. Therefore, No. 5 is an optimal composition.

V. CONCLUSION

In this paper, we introduce a model named Swin-MIQT, designed to blindly assess image quality. Swin-MIQT leverages a transformer-based architecture to learn multi-scale representations, enabling it to process images at their original resolution, similar to typical vision transformers. Our results demonstrate that the multi-scale representations effectively capture quality information across various granularities, thereby enhancing model performance. Swin-MIQT exhibits robust performance in handling both synthetic and authentic distortions, achieving state-of-the-art (SOTA) results on three synthetically distorted databases and competitive performance on three authentically distorted databases. Finally, Swin-MIQT is notable for its extensibility, featuring five modules that can be seamlessly upgraded with alternative architectures, potentially enhancing model performance even further.

REFERENCES

[1] W. Zhang, K. Ma, J. Yan, D. Deng, and Z. Wang, "Blind image quality assessment using a deep bilinear convolutional neural network," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 1, pp. 36–47, 2020.

[2] D. Varga, "No-reference image quality assessment based on the fusion of statistical and perceptual features," *Journal of Imaging*, vol. 6, no. 8, p. 75, 2020.

[3] J. Galbally, S. Marcel, and J. Fierrez, "Image quality assessment for fake biometric detection: Application to iris, fingerprint, and face recognition," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 710–724, 2013.

[4] M. Loubele, R. Jacobs, F. Maes, F. Schutyser, D. Debaveye, R. Bogaerts, W. Coudyzer, D. Vandermeulen, J. Van Cleynenbreugel, G. Marchal *et al.*, "Radiation dose vs. image quality for low-dose ct protocols of the head for maxillofacial surgery and oral implant planning," *Radiation Protection Dosimetry*, vol. 117, no. 1-3, pp. 211–216, 2005.

[5] J. San Pedro and S. Siersdorfer, "Ranking and classifying attractiveness of photos in folksonomies," in *Proceedings of the 18th International Conference on World Wide Web*, 2009, pp. 771–780.

[6] L. Zhang, L. Zhang, and A. C. Bovik, "A feature-enriched completely blind image quality evaluator," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2579–2591, 2015.

[7] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Transactions on Image Processing*, vol. 25, no. 9, pp. 4444–4457, 2016.

[8] J. Kim and S. Lee, "Fully deep blind image quality predictor," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 206–220, 2016.

[9] D. Ghadiyaram and A. C. Bovik, "Perceptual quality prediction on authentically distorted images using a bag of features approach," *Journal of Vision*, vol. 17, no. 1, pp. 32–32, 2017.

[10] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, and W. Zuo, "End-to-end blind image quality assessment using deep neural networks," *IEEE Transactions on Image Processing*, vol. 27, no. 3, pp. 1202–1213, 2017.

[11] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "Metaiqa: Deep meta-learning for no-reference image quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 14 143–14 152.

[12] W. Kim, A.-D. Nguyen, S. Lee, and A. C. Bovik, "Dynamic receptive field generation for full-reference image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4219–4231, 2020.

[13] Y. Liu, G. Zhai, K. Gu, X. Liu, D. Zhao, and W. Gao, "Reduced-reference image quality assessment in free-energy principle and sparse representation," *IEEE Transactions on Multimedia*, vol. 20, no. 2, pp. 379–391, 2017.

[14] Y. Liu, K. Gu, Y. Zhang, X. Li, G. Zhai, D. Zhao, and W. Gao, "Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 30, no. 4, pp. 929–943, 2019.

[15] P. Le Callet and F. Autrusseau, "Subjective quality assessment irccyn/ivc database," 2005. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00580755

[16] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," *IEEE Transactions on Image Processing*, vol. 15, no. 11, pp. 3440–3451, 2006.

[17] E. C. Larson and D. M. Chandler, "Most apparent distortion: full-reference image quality assessment and the role of strategy," *Journal of Electronic Imaging*, vol. 19, no. 1, p. 011006, 2010.

[18] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020.

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

[20] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "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*, 2020, pp. 3575–3585.

[21] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *2012 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, 2012, pp. 1098–1105.

[22] A. Mittal, A. K. Moorthy, and A. C. Bovik, "No-reference image quality assessment in the spatial domain," *IEEE Transactions on Image Processing*, vol. 21, no. 12, pp. 4695–4708, 2012.

[23] S. Bianco, L. Celona, P. Napoletano, and R. Schettini, "On the use of deep learning for blind image quality assessment," *Signal, Image and Video Processing*, vol. 12, no. 2, pp. 355–362, 2018.

[24] D. Varga, D. Saupe, and T. Szirányi, "Deeprn: A content preserving deep architecture for blind image quality assessment," in *2018 IEEE International Conference on Multimedia and Expo (ICME)*. IEEE, 2018, pp. 1–6.

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

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

[27] F. Gao, D. Tao, X. Gao, and X. Li, "Learning to rank for blind image quality assessment," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 10, pp. 2275–2290, 2015.

[28] K. Ma, W. Liu, T. Liu, Z. Wang, and D. Tao, "dipiq: Blind image quality assessment by learning-to-rank discriminable image pairs," *IEEE Transactions on Image Processing*, vol. 26, no. 8, pp. 3951–3964, 2017.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3531416

[29] X. Liu, J. Van De Weijer, and A. D. Bagdanov, "Rankiqa: Learning from rankings for no-reference image quality assessment," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1040–1049.

[30] K.-Y. Lin and G. Wang, "Hallucinated-iqa: No-reference image quality assessment via adversarial learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 732–741.

[31] Y. Ma, X. Cai, F. Sun, and S. Hao, "No-reference image quality assessment based on multi-task generative adversarial network," *IEEE Access*, vol. 7, pp. 146 893–146 902, 2019.

[32] J. You and J. Korhonen, "Transformer for image quality assessment," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 1389–1393.

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

[34] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *Advances in Neural Information Processing Systems*, vol. 30, 2017. [Online]. Available: https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

[35] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[36] ——, "Spatial pyramid pooling in deep convolutional networks for visual recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 9, pp. 1904–1916, 2015.

[37] Z. Liu, Y. Lin, Y. Cao, H. Hu, Y. Wei, Z. Zhang, S. Lin, and B. Guo, "Swin transformer: Hierarchical vision transformer using shifted windows," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 012–10 022.

[38] D. Jayaraman, A. Mittal, A. K. Moorthy, and A. C. Bovik, "Objective quality assessment of multiply distorted images," in *2012 Conference Record of the Forty Sixth Asilomar Conference on Signals, Systems and Computers (ASILOMAR)*. IEEE, 2012, pp. 1693–1697.

[39] N. Ponomarenko, L. Jin, O. Ieremeiev, V. Lukin, K. Egiazarian, J. Astola, B. Vozel, K. Chehdi, M. Carli, F. Battisti *et al.*, "Image database tid2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, 2015.

[40] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, 2015.

[41] S. Su, Q. Yan, Y. Zhu, C. Zhang, X. Ge, J. Sun, and Y. Zhang, "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*, 2020, pp. 3667–3676.

[42] S. Yang, T. Wu, S. Shi, S. Lao, Y. Gong, M. Cao, J. Wang, and Y. Yang, "Maniqa: Multi-dimension attention network for no-reference image quality assessment," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022, pp. 1191–1200.

[43] N. Carion, F. Massa, G. Synnaeve, N. Usunier, A. Kirillov, and S. Zagoruyko, "End-to-end object detection with transformers," in *European Conference on Computer Vision*. Springer, 2020, pp. 213–229.

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

[45] H. Touvron, M. Cord, M. Douze, F. Massa, A. Sablayrolles, and H. Jégou, "Training data-efficient image transformers & distillation through attention," in *International Conference on Machine Learning*. PMLR, 2021, pp. 10 347–10 357. [Online]. Available: https://proceedings.mlr.press/v139/touvron21a

[46] M. Cheon, S.-J. Yoon, B. Kang, and J. Lee, "Perceptual image quality assessment with transformers," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2021, pp. 433–442.

[47] J. Gu, H. Cai, C. Dong, J. S. Ren, Y. Qiao, S. Gu, R. Timofte, M. Cheon, S. Yoon, B. K. Kang, J. Lee, Q. Zhang, H. Guo, Y. Bin, Y. Hou, H. Luo, J. Guo, Z. Wang, H. Wang, W. Yang, Q. Bai, S. Shi, W. Xia, M. Cao, J. Wang, Y. Chen, Y. Yang, Y. Li, T. Zhang, L. Feng, Y. Liao, J. Li, W. Thong, J. C. Pereira, A. Leonardis, S. McDonagh, K. Xu, L. Yang, H. Cai, P. Sun, S. M. Ayyoubzadeh, A. Royat, S. A. Fezza, D. Hammou, W. Hamidouche, S. Ahn, G. Yoon, K. Tsubota, H. Akutsu, and K. Aizawa, "Ntire 2021 challenge on perceptual image quality assessment," in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2021, pp. 677–690.

[48] G. Jinjin, C. Haoming, C. Haoyu, Y. Xiaoxing, J. S. Ren, and D. Chao, "Pipal: a large-scale image quality assessment dataset for perceptual image restoration," in *European Conference on Computer Vision*. Springer, 2020, pp. 633–651.

[49] C. Szegedy, S. Ioffe, V. Vanhoucke, and A. A. Alemi, "Inception-v4, inception-resnet and the impact of residual connections on learning," in *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. [Online]. Available: https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/viewPaper/14806

[50] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *2009 IEEE Conference on Computer Vision and Pattern Recognition*. Ieee, 2009, pp. 248–255.

[51] D. Hendrycks and K. Gimpel, "Gaussian error linear units (gelus)," *arXiv preprint arXiv:1606.08415*, 2016.

[52] H. Talebi and P. Milanfar, "Nima: Neural image assessment," *IEEE Transactions on Image Processing*, vol. 27, no. 8, pp. 3998–4011, 2018.

[53] W. Xue, X. Mou, L. Zhang, A. C. Bovik, and X. Feng, "Blind image quality assessment using joint statistics of gradient magnitude and laplacian features," *IEEE Transactions on Image Processing*, vol. 23, no. 11, pp. 4850–4862, 2014.

[54] A. K. Moorthy and A. C. Bovik, "Blind image quality assessment: From natural scene statistics to perceptual quality," *IEEE Transactions on Image Processing*, vol. 20, no. 12, pp. 3350–3364, 2011.

[55] W. Xue, L. Zhang, and X. Mou, "Learning without human scores for blind image quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 995–1002.

[56] X. Yang, F. Li, and H. Liu, "A survey of dnn methods for blind image quality assessment," *IEEE Access*, vol. 7, pp. 123 788–123 806, 2019.

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

[58] J. Liang, J. Cao, G. Sun, K. Zhang, L. Van Gool, and R. Timofte, "Swinir: Image restoration using swin transformer," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 1833–1844.

GENG FU received the B.S. degree in computer science and technology from Qilu Normal University, Jinan, China, in 2021. He is currently pursuing the M.S. degree in electronic information at Shandong Univrsity of Finance and Economics, Jinan, China. His main research interests include computer vision and medical imaging.

ZIYU WANG graduated from Shandong Medical College. He is currently a senior technician in Yidu Central Hospital of Weifang. His main research interests include image processing and medical imaging.

CUIJUAN ZHANG graduated from Weifang Nursing Vocational College. She is currently a supervisor nurse in Yidu Central Hospital of Weifang. Her main research interests include image processing and medical imaging.

ZERONG QI received the B.S. and M.S. degrees from Shandong Agricultural University and Beijing Jiaotong University, respectively. He is currently pursuing the PH.D. degree in computational mathematics at Shandong University. He is a technical supervisor in Shandong Chengshi Electronic Technology Limited Company. His main research interests include computer vision, internet of things, public safety technology and medical imaging.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2025.3531416

MINGZHENG HU received the B.S. and M.S. degrees from Shandong University. He is a technical supervisor in Shandong Chengshi Electronic Technology Limited Company. His main research interests include computer vision, internet of things, public safety technology and medical imaging.

SHUJUN FU received the B.S. and M.S. degrees from Shandong University, PH.D. degree from Beijing Jiaotong University, respectively. He is currently a professor in Shandong University. He has published more than one hundred academic papers in computer vision, inverse problem and medical imaging.

YUNFENG ZHANG received the PH.D. degree from Shandong University. He is currently a professor in Shandong Univrsity of Finance and Economics. His main research interests include computer vision, digital financial analysis and medical imaging.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/