MAMIQA: No-Reference Image Quality Assessment Based on Multiscale Attention Mechanism With Natural Scene Statistics

Li Yu, Junyang Li, Farhad Pakdaman, *Member, IEEE*, Miaogen Ling, and Moncef Gabbouj, *Fellow, IEEE*

***Abstract*—No-Reference Image Quality Assessment aims to evaluate the perceptual quality of an image, according to human perception. Many recent studies use Transformers to assign different self-attention mechanisms to distinguish regions of an image, simulating the perception of the human visual system (HVS). However, the quadratic computational complexity caused by the self-attention mechanism is time-consuming and expensive. Meanwhile, the image resizing in the feature extraction stage loses the full-size image quality. To address these issues, we propose a lightweight attention mechanism using decomposed large-kernel convolutions to extract multiscale features, and a novel feature enhancement module to simulate HVS. We also propose to compensate the information loss caused by image resizing, with supplementary features from natural scene statistics. Experimental results on five standard datasets show that the proposed method surpasses the SOTA, while significantly reducing the computational costs.**

***Index Terms*—Human visual system (HVS), large kernel attention, multiscale feature extraction, NR-IQA.**

I. Introduction

PRISTINE quality images are subject to varying degrees of degradation during processing such as acquisition, compression, transmission, and storage, resulting in loss of visual information. Objective image quality evaluation has an essential role in image signal processing [1], [2]. Among all Image Quality Assessment (IQA) paradigms, no-reference image quality assessment (NR-IQA) has become a research hotspot because of its wide applications, especially when the reference image is not available.

Manuscript received 10 March 2023; revised 9 May 2023; accepted 9 May 2023. Date of publication 16 May 2023; date of current version 25 May 2023. This work was supported in part by the National Natural Science Foundation of China under Grant 62002172, in part by the European Union's Horizon 2020 under the Marie Skłodowska-Curie under Grant 101022466, and in part by the Business Finland AMALIA-2023 under Grant 97/31/2023 The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Demetrio Labate. *(Corresponding author: Moncef Gabbouj.)*

Li Yu, Junyang Li, and Miaogen Ling are with the School of Computer Science, Nanjing University of Information Science and Technology, Nanjing 210044, China, and also with the Engineering Research Center of Digital Forensics Ministry of Education, Nanjing University of Information Science and Technology, Nanjing 210044, China (e-mail: mailofyuli@126.com; 20211249427@nuist.edu.cn; mgling@nuist.edu.cn).

Farhad Pakdaman and Moncef Gabbouj are with the Faculty of Information Technology and Communication Sciences, Tampere University, 33101 Tampere, Finland (e-mail: farhad.pakdaman@tuni.fi; moncef.gabbouj@tuni.fi).

Digital Object Identifier 10.1109/LSP.2023.3276645

With the development of deep learning, several NR-IQA methods[3], [4], [5], [6], [7], [8] based on Convolutional Neural Networks (CNNs) have been proposed. The main problem with these methods is that only fixed-sized image inputs are allowed, which leads to image resizing as preprocessing for some inputs. For the IQA task, this will affect the quality of the image and leads to deviations in the final prediction.

Neuroscience research [9] has demonstrated that when the human visual system (HVS) estimates the quality of an image, some regions are given higher attention. Therefore, HVS can be simulated by introducing the attention mechanism to assign different weights to different regions of the image. Transformers have been applied to many machine vision tasks [10], [11], [12], [13], [14], [15]. Several NR-IQA methods also use transformers to implement the attention mechanism of HVS. TRIQ [16] uses convolutional neural networks to extract features and implements a self-attention mechanism with the Transformer's encoder, capable of handling image inputs of different resolutions. TranSLA [17] introduces saliency information to guide the self-attention mechanism of the Transformer while incorporating a gradient map to supplement the Transformer with local information. TReS [18] employs a hybrid approach of CNN and a self-attention mechanism in Transformer to extract local and non-local features from the input image. However, recent research [19] argues that the computation of a one-dimensional structure for a two-dimensional structured image is unreasonable and poses many challenges. First, the self-attention mechanism imposes a quadratic computational complexity for images, which is time-consuming and expensive for high-resolution images. Such solutions with high computational complexity and storage requirements cannot be easily deployed in mobile or edge devices [20]. Second, Transformer-based models [16], [17], [18] usually use CNN as the feature extractor, which still performs a resize operation on the image; hence, deviating from the original image quality.

To solve the above-mentioned issues, we propose a new NR-IQA method based on a multiscale attention mechanism called MAMIQA, which is divided into two branches: one branch mimics the HVS and the other branch captures original image features to compensate for the information loss due to image resizing. For the first branch, we use decomposed large kernel convolution to assign different attentions to different regions of the image. By decoupling the large kernel convolution into

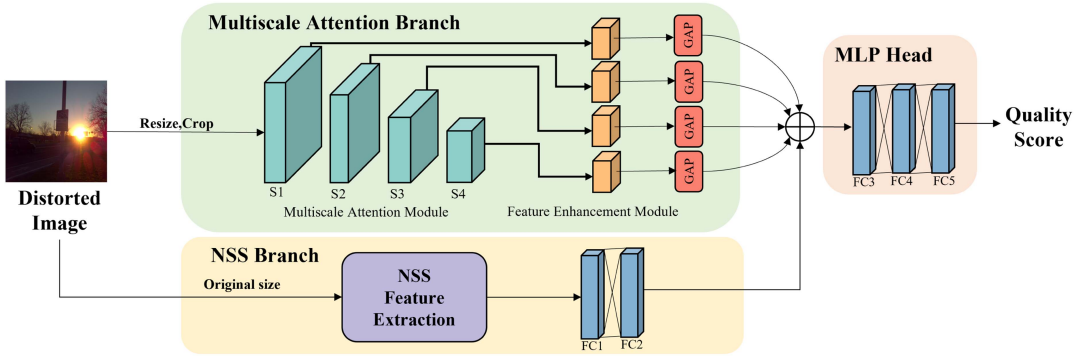

Fig. 1. Framework of the proposed MAMIQA. (a) In the multiscale attention branch, the image is first fed into the multiscale attention module, then features are enhanced in feature enhancement module. (b) In the NSS branch, the features of the original size image are extracted, and then sent to a two-layer fully connected layer. (c) Finally, the features extracted from the two branches are sent to MLP head for feature fusion and prediction of the final quality score.

depth-wise convolution, depth-wise dilation convolution and channel convolution, we implement the attention mechanism with a lower computational overhead than Transformers. Moreover, the HVS views images at multiple scales [21]. To simulate the behavior of the HVS, we extract features from multiple scales and further propose a feature enhancement module (FEM) to enrich the local fine-grained details and global semantic information of multi-scale features. For the second branch, the traditional natural scene statistics (NSS) [22] method is used. The advantage of NSS methods [23], [24], [25] is that there is no limitation of input image size, i.e., the original image features can be extracted without resizing or cropping the image. Therefore, NSS is adopted to extract the original image features, to compensate for the information loss due to image resizing. The main contributions of this letter are as follows:

- We propose a novel NR-IQA method based on light-weight attention mechanism, which mimics the HVS and improves the performance of image quality assessment.
- We adopt natural scene statistics to extract the features of original-sized images, which compensates for the information loss caused by image resizing or cropping.
- We extract multiscale features and propose an efficient feature enhancement module (FEM) to improve the multiscale feature representation of the model.
- We enable low complexity IQA, by enhancing the performance beyond transformers-based methods, with much lower computational complexity.

II. Proposed Method

Fig. 1 depicts the overall framework of our proposed MAMIQA, which is a two-branch network structure with a multiscale attention branch, and an NSS branch.

For the multiscale attention branch, we extract attention features of four scales through the multiscale attention module. The extracted four-scale features are then fed into our proposed feature enhancement module (FEM) for feature enhancement. Finally, global average pooling (GAP) is performed to obtain the enhanced multiscale features.

For the NSS branch, we use BRISQUE-based NSS features [24]. First, we calculate the mean subtracted contrast normalized (MSCN) coefficients of the image, i.e. the local normalized luminance coefficients. Since the MSCN coefficients of the original quality image conform to the Gaussian distribution, while the distorted image does not conform to this statistical regulation [22], we can extract the features of the image by quantifying the difference between the two. We capture this regular deviation by using generalized Gaussian distribution (GGD). Also, the product of the MSCN adjacent coefficients of the image is shown to conform to the statistical regularity of the natural image. Hence, the deviation of the product coefficients is captured by using an asymmetric generalized Gaussian distribution (AGGD). The NSS features are extracted from both 1 and 1/2 scales, and fed into two fully connected layers, which use PReLu as the activation function (as PReLu can strengthen the nonlinear relationship between the layers and thus, accelerate the training process).

We use a three-layer fully connected MLP Head to fuse the above features and predict the perceptual quality. For each batch of images in training, the regression loss is minimized with the mean absolute error (MAE) as the loss function, for stable training and preventing gradient explosion [26].

$$L = \frac{1}{N}\sum_{i}^{N} \|q_i - s_i\| \tag{1}$$

Here, q_i is the predicted quality score for i_{th} image and s_i is its corresponding subjective quality score (ground truth).

A. Multiscale Attention Module (MAM)

The key to the attentional mechanism is to produce the attention map, which shows the importance of the different regions, so we should learn the dependency between different regions. There are two approaches to capture such dependencies. One approach is implemented through the self-attention mechanism, but with obvious drawbacks, which have been listed above. The other approach is to establish correlations using large kernel convolutions. Implementing attention directly using large kernel convolutions leads to high computational overheads as well as a large number of parameters. To solve this problem, we apply an attention network [19] as the backbone of the multiscale attention module. A large kernel convolution is decomposed to

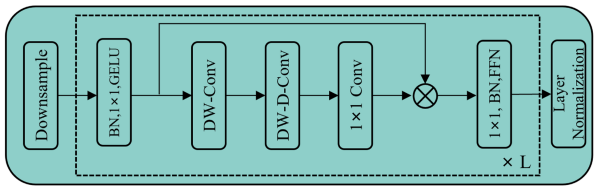

Fig. 2. A stage of Multiscale Attention Module.

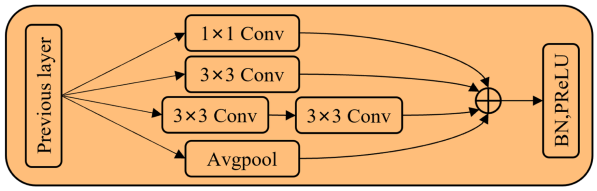

Fig. 3. The proposed feature enhancement module (FEM).

capture the dependencies among different regions. As shown in Fig. 2, a large kernel convolution can be divided into three components: a depth-wise convolution, a depth-wise dilation convolution, and a channel convolution. This module captures the feature dependencies with a lower computational cost and lower number of parameters. The input features are first downsampled to obtain F_1, then L groups of operations are stacked in sequence to extract the features. The operations of each group can be described as follows:

$$\begin{cases} F_2 = f_{\text{GELU}}\left(f_{1\times1}\left(f_{BN}\left(F_1\right)\right)\right) \\ Attention = f_{1\times1}\left(f_{DW\_D\_Conv}\left(f_{DW\_Conv}\left(F_2\right)\right)\right) \\ F_3 = F_2 \otimes Attention \\ F_4 = f_{FFN}\left(f_{BN}\left(f_{1\times1}\left(F_3\right)\right)\right) \end{cases} \tag{2}$$

where f_{GELU} denotes GELU activation, $f_{BN}(\cdot)$ denotes batch normalization, $f_{1\times1}(\cdot)$ denotes 1×1 convolution, $f_{DW\_Conv}(\cdot)$ and $f_{DW\_D\_Conv}(\cdot)$ denote depth-wise convolution and depth-wise dilation convolution, respectively. $f_{FFN}(\cdot)$ denotes the convolutional feed-forward network, $\otimes$ denotes element-wise product, L for each stage are $\{3, 3, 12, 3\}$. Finally, the layer normalization is applied at the end of each stage.

B. Feature Enhancement Module (FEM)

The human visual system views images at multiple scales [21]. To help the network better understand the content information in distorted images, four scales of features are extracted with the multiscale attention module, then enhanced, and finally processed by global average pooling to stitch together the multiscale features. In order to enhance the local and global feature representation, inspired by the inception module in [40], we propose a feature enhancement module (FEM). As shown in Fig. 3, the FEM consists of 1×1, 3×3 and 5×5 convolutional layers with different reception fields and an average pooling layer in parallel, where convolutional kernels of different sizes can extract information of different scales. Using average pooling can reduce the dimension of features, remove redundant information, and fuse multi-dimensional features to extract more dense features. In addition, we use two 3×3 convolutional kernels instead of one 5×5 convolutional kernel. We use two different stride 3×3 convolution layers to achieve the same reception field as the 5×5 convolution layer, which has the advantage of having fewer parameters and reducing the complexity of the network.

III. Experiments

For the multiscale attention branch, the distorted images are randomly cropped into 224×224 patches. To augment the training samples, a random horizontal flip with a vertical flip is performed. For the NSS branch, we use the original size image as input. The proposed MAMIQA is implemented with Pytorch and trained under Ubuntu 16.04 operation system with TITAN RTX GPU. We use the Adam optimizer with weight decay 5×10^{-4} to train our model. The batch size is set to 64, and the learning rate is set to 2×10^{-5}. To validate the performance of the proposed method, we conducted experiments on five publicly available IQA Datasets, including three synthetically distorted datasets LIVE II [41], TID2013 [42], CSIQ [43], and two authentically distorted datasets LIVE-C [44], KonIQ-10K [45]. Two widely used metrics, Spearman rank-order correlation coefficient (SRCC) and Pearson linear correlation coefficient (PLCC), are adopted to measure the performance of IQA models against the ground truth subjective quality.

A. Performance Evaluation

Table I shows the performance of our proposed method compared to other methods on five IQA datasets, with first three lines indicate the FR-IQA methods and the remaining NR-IQA methods. The experimental results of the competing methods are based on implementations obtained from the original papers. Our proposed model achieves superior performance in PLCC and SRCC. Specifically, our model achieves 4.2%, 3.8% (PLCC, SRCC) higher than MS-GMSD (best FR-IQA) in TID2013 dataset and 1.8%, 1.7% (PLCC, SRCC) higher than MS-GMSD in LIVE dataset. Compared with other NR-IQA models, our model gains 1.3%, 1.6% (PLCC, SRCC) in CSIQ over DBCNN (second-best).

Table II reports the experiments conducted over the datasets to further compare our method with SOTA methods. All methods were trained on one dataset and tested on three other datasets without any fine-tuning or parameter adjustment. It is observed that the proposed method achieved the best results in 8 out of the 12 tested cases, and competitive results in the remaining 4 cases, showcasing the strong generalization ability of the proposed method.

We further compare the complexity of our model with two FR-IQA methods PieAPP and DISTS, and three NR-IQA methods (including CNN-based method (ResNet-50), Transformer-based methods (TReS) and StairIQA). As can be seen from Table IV, our method outperforms FR-IQA methods with a large margin, while requiring significantly lower computations. DISTS has the lowest number of parameters among the competing methods. However, MAMIQA makes up for this by a better performance with a much lower computation load. For ResNet50, it replaced the MAM module in the proposed method. The experimental results show that the proposed model requires only half the computation of the ResNet-50 model (in terms of Giga-Flops), and only one-fifth of the number of parameters, while achieving

TABLE I
COMPARISON OF OUR PROPOSED METHOD WITH SOTA ALGORITHMS ON THREE SYNTHETICALLY DISTORTION DATASETS AND TWO AUTHENTICALLY DISTORTION DATASETS. THE TOP TWO RESULTS ARE SHOWN IN BOLD

	LIVE II		CSIQ		TID2013		LIVEC		KonIQ	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
MS-GMSD [27]	0.963	0.964	0.951	**0.954**	0.895	0.890	/	/	/	/
PieAPP [28]	0.908	0.919	0.842	0.907	0.836	0.875	/	/	/	/
DISTS [29]	0.954	0.954	0.928	0.929	0.855	0.830	/	/	/	/
DIIVINE [30]	0.908	0.892	0.776	0.804	0.567	0.643	0.591	0.588	0.558	0.546
BRISQUE [24]	0.944	0.929	0.748	0.812	0.571	0.626	0.629	0.629	0.685	0.681
BPRI [31]	0.930	0.929	0.918	0.896	0.890	0.899	/	/	/	/
BMPRI [32]	0.933	0.931	0.934	0.909	**0.947**	**0.929**	/	/	/	/
WaDIQaM [33]	0.955	0.960	0.844	0.852	0.855	0.835	0.671	0.682	0.807	0.804
MEON [34]	0.955	0.951	0.864	0.852	0.824	0.808	0.710	0.697	0.628	0.611
DBCNN [5]	0.971	0.968	**0.959**	0.946	0.865	0.816	0.869	0.869	0.884	0.875
Hall-IQA [35]	**0.978**	**0.976**	0.906	0.892	0.846	0.834	/	/	/	/
MetaIQA [36]	0.959	0.960	0.908	0.899	0.868	0.856	0.802	0.835	0.856	0.887
HyperIQA [37]	0.966	0.962	0.942	0.923	0.858	0.840	0.882	0.859	0.917	0.906
TIQA [16]	0.965	0.949	0.838	0.825	0.858	0.846	0.861	0.845	0.903	0.892
TReS [18]	0.968	0.969	0.942	0.922	0.883	0.863	0.877	0.846	0.928	0.915
VCRNet [38]	0.974	0.973	0.955	0.943	0.875	0.846	0.865	0.856	0.909	0.894
StairIQA [39]	0.970	0.966	0.941	0.919	/	/	**0.918**	**0.899**	**0.936**	**0.921**
Proposed	**0.981**	**0.981**	**0.972**	**0.962**	**0.937**	**0.928**	**0.895**	**0.874**	**0.937**	**0.926**

TABLE II
SRCC RESULTS FOR CROSS-DATASET EXPERIMENTS

Train on	LIVE II			TID2013			CSIQ			LIVEC		
Test on	TID2013	CSIQ	LIVEC	LIVE	CSIQ	LIVEC	LIVE II	TID2013	LIVEC	LIVE II	TID2013	CSIQ
DIIVINE [30]	0.342	0.602	0.296	0.714	0.583	0.235	0.817	0.417	0.366	0.354	0.327	0.419
BRISQUE [24]	0.354	0.573	0.326	0.724	0.568	0.109	0.823	0.433	0.106	0.244	0.275	0.236
WaDIQaM [33]	0.396	0.601	0.151	0.805	0.683	0.009	0.813	0.506	0.106	0.323	0.141	0.323
DBCNN [5]	0.536	0.762	0.552	0.872	0.703	0.412	0.871	0.523	0.453	**0.757**	0.401	0.631
Hall-IQA [35]	0.486	0.668	0.126	0.786	0.683	0.116	0.833	0.491	0.107	/	/	/
VCRNet [38]	0.502	**0.768**	**0.615**	0.822	0.721	0.307	0.886	0.542	0.463	0.746	**0.416**	0.566
Proposed	**0.587**	0.753	0.559	**0.885**	**0.785**	**0.376**	**0.934**	**0.566**	**0.477**	0.713	0.386	**0.674**

TABLE III
ABLATION EXPERIMENTS ON DIFFERENT COMPONENTS

	LIVE II		CSIQ		TID2013		LIVEC		KonIQ	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
multiscale attention module (MAM)	0.961	0.976	0.956	0.948	0.901	0.884	0.875	0.850	0.929	0.920
MAM+FEM	**0.982**	**0.981**	0.963	0.949	0.910	0.895	0.887	0.861	0.936	**0.926**
MAM+FEM+NSS (proposed)	0.981	**0.981**	**0.972**	**0.962**	**0.937**	**0.928**	**0.895**	**0.874**	**0.937**	**0.926**

TABLE IV
COMPARISON OF SRCC RESULTS AND COMPLEXITY WITH COMPETITIVE METHODS

Method	GFlops	Params.(M)	CSIQ	TID2013	Koniq
PieAPP	34.03	68.38	0.907	0.875	/
DISTS	30.69	14.72	0.929	0.830	/
Resnet50+FEM+NSS	10.35	189.71	0.913	0.905	0.896
TReS	8.39	34.46	0.922	0.863	0.915
StairIQA	10.38	31.80	0.919	/	0.921
MAMIQA(proposed)	5.45	38.68	0.962	0.928	0.926

improved performance. When compared to the Transformer-based model (TReS), the proposed method has roughly the same number of parameters, but with 35% lower computational complexity and much higher performance. This validates the efficiency of the proposed lightweight solution.

B. Ablation Study

In order to analyze the effectiveness of using the feature enhancement module and NSS methods, an ablation experiment was conducted to verify the influence of each component in the proposed model. We constructed the following model settings: 1) a model containing only the multiscale attention module backbone network. 2) A model containing a multiscale attention module network with a feature enhancement module (FEM). 3) The proposed model. Table III reports the results for the LIVE, CSIQ, TID2013, LIVEC, and KONIQ datasets. The results show how adding each of the FEM and NSS improve the performance, for both synthetic and authentic distortions, indicating the effectiveness of the proposed modules.

IV. CONCLUSION

In this letter we proposed a lightweight IQA method named MAMIQA. A multiscale attention branch captures the attention via decomposed large kernel operations and enhances the feature representation via a novel feature enhancement module. An NSS branch is used to extract supplementary features to compensate the information loss caused by image cropping. The experimental results showed that the proposed model gains a significant performance improvement over the SOTA. Specifically, it was demonstrated that the proposed method better estimates the quality compared to the trending Transformers-based methods, while requiring significantly lower computational power.

ACKNOWLEDGMENT

The authors acknowledge the High Performance Computing Center of Nanjing University of Information Science & Technology for their support of this work.

REFERENCES

[1] G. Zhai and X. Min, "Perceptual image quality assessment: A survey," *Sci. China Inf. Sci.*, vol. 63, pp. 1–52, 2020.

[2] X. Min et al., "Screen content quality assessment: Overview, benchmark, and beyond," *ACM Comput. Surv.*, vol. 54, no. 9, pp. 1–36, 2021.

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

[4] J. Kim, A.-D. Nguyen, and S. Lee, "Deep CNN-based blind image quality predictor," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 1, pp. 11–24, Jan. 2019.

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

[6] P. Chen, L. Li, Q. Wu, and J. Wu, "SPIQ: A self-supervised pre-trained model for image quality assessment," *IEEE Signal Process. Lett.*, vol. 29, pp. 513–517, 2022.

[7] S. V. R. Dendi, C. Dev, N. Kothari, and S. S. Channappayya, "Generating image distortion maps using convolutional autoencoders with application to no reference image quality assessment," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 89–93, Jan. 2019.

[8] Q. Jiang, Z. Peng, S. Yang, and F. Shao, "Authentically distorted image quality assessment by learning from empirical score distributions," *IEEE Signal Process. Lett.*, vol. 26, no. 12, pp. 1867–1871, Dec. 2019.

[9] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?," *Nature Rev. Neurosci.*, vol. 5, no. 6, pp. 495–501, 2004.

[10] J. Kong, Y. Bian, and M. Jiang, "MTT: Multi-scale temporal transformer for skeleton-based action recognition," *IEEE Signal Process. Lett.*, vol. 29, pp. 528–532, 2022.

[11] Z. Liang, Y. Wang, L. Wang, J. Yang, and S. Zhou, "Light field image super-resolution with transformers," *IEEE Signal Process. Lett.*, vol. 29, pp. 563–567, 2022.

[12] T. Li, Z. Zhang, L. Pei, and Y. Gan, "HashFormer: Vision transformer based deep hashing for image retrieval," *IEEE Signal Process. Lett.*, vol. 29, pp. 827–831, 2022.

[13] M. Zhao, G. Cao, X. Huang, and L. Yang, "Hybrid transformer-CNN for real image denoising," *IEEE Signal Process. Lett.*, vol. 29, pp. 1252–1256, 2022.

[14] H. Yin and S. Ma, "CSformer: Cross-scale features fusion based transformer for image denoising," *IEEE Signal Process. Lett.*, vol. 29, pp. 1809–1813, 2022.

[15] H. Wu, C. Xu, and H. Liu, "SMART: Semantic-aware masked attention relational transformer for multi-label image recognition," *IEEE Signal Process. Lett.*, vol. 29, pp. 2158–2162, 2022.

[16] J. You and J. Korhonen, "Transformer for image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, 2021, pp. 1389–1393.

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

[18] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-reference image quality assessment via transformers, relative ranking, and self-consistency," in *Proc. IEEE/CVF Winter Conf. Appl. Comput. Vis.*, 2022, pp. 1220–1230.

[19] M.-H. Guo, C.-Z. Lu, Z.-N. Liu, M.-M. Cheng, and S.-M. Hu, "Visual attention network," 2022, *arXiv:2202.09741*.

[20] Z. Mei, Y.-C. Wang, X. He, and C.-C. J. Kuo, "GreenBIQA: A lightweight-blind image quality assessment method," in *Proc. IEEE 24th Int. Workshop Multimedia Signal Process.*, 2022, pp. 1–6.

[21] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing," *RCA Engineer*, vol. 29, no. 6, pp. 33–41, 1984.

[22] D. L. Ruderman, "The statistics of natural images," *Netw.: Computation Neural Syst.*, vol. 5, no. 4, pp. 517–548, 1994.

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

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

[25] A. Mittal, R. Soundararajan, and A. C. Bovik, "Making a "completely blind" image quality analyzer," *IEEE Signal Process. Lett.*, vol. 20, no. 3, pp. 209–212, Mar. 2013.

[26] J. Qi, J. Du, S. M. Siniscalchi, X. Ma, and C.-H. Lee, "On mean absolute error for deep neural network based vector-to-vector regression," *IEEE Signal Process. Lett.*, vol. 27, pp. 1485–1489, 2020.

[27] B. Zhang, P. V. Sander, and A. Bermak, "Gradient magnitude similarity deviation on multiple scales for color image quality assessment," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, 2017, pp. 1253–1257.

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

[29] K. Ding, K. Ma, S. Wang, and E. P. Simoncelli, "Image quality assessment: Unifying structure and texture similarity," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 5, pp. 2567–2581, May 2022.

[30] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind image quality assessment: A natural scene statistics approach in the DCT domain," *IEEE Trans. Image Process.*, vol. 21, no. 8, pp. 3339–3352, Aug. 2012.

[31] X. Min, K. Gu, G. Zhai, J. Liu, X. Yang, and C. W. Chen, "Blind quality assessment based on pseudo-reference image," *IEEE Trans. Multimedia*, vol. 20, no. 8, pp. 2049–2062, Aug. 2018.

[32] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 508–517, Jun. 2018.

[33] S. Bosse, D. Maniry, K.-R. Müller, T. Wiegand, and W. Samek, "Deep neural networks for no-reference and full-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 206–219, Jan. 2018.

[34] 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 Trans. Image Process.*, vol. 27, no. 3, pp. 1202–1213, Mar. 2018.

[35] K.-Y. Lin and G. Wang, "Hallucinated-IQA: No-reference image quality assessment via adversarial learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 732–741.

[36] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep meta-learning for no-reference image quality assessment," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 14143–14152.

[37] S. Su et al., "Blindly assess image quality in the wild guided by a self-adaptive hyper network," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 3664–3673.

[38] Z. Pan, F. Yuan, J. Lei, Y. Fang, X. Shao, and S. Kwong, "VCRNet: Visual compensation restoration network for no-reference image quality assessment," *IEEE Trans. Image Process.*, vol. 31, pp. 1613–1627, 2022.

[39] W. Sun, X. Min, D. Tu, S. Ma, and G. Zhai, "Blind quality assessment for in-the-wild images via hierarchical feature fusion and iterative mixed database training," *IEEE J. Sel. Topics Signal Process.*, early access, Apr. 26, 2023, doi: 10.1109/JSTSP.2023.3270621.

[40] C. Szegedy et al., "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.

[41] H. Sheikh, "Live image quality assessment database release 2," 2005. [Online]. Available: http://live.ece.utexas.edu/research/quality

[42] N. Ponomarenko et al., "Color image database tid2013: Peculiarities and preliminary results," in *Proc. IEEE Eur. Workshop Vis. Inf. Process.*, 2013, pp. 106–111.

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

[44] D. Ghadiyaram and A. C. Bovik, "Massive online crowdsourced study of subjective and objective picture quality," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 372–387, Jan. 2016.

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