BlockIQA: Local Sensitivity-Enhanced Blind Image Quality Assessment through Deep Block Analysis

Yuqi Pang
South China University of Technology
School of Electronic and Information Engineering
Guangzhou, Guangdong, China
eepangyuqi@mail.scut.edu.cn

Yican Liu
South China University of Technology
School of Electronic and Information Engineering
Guangzhou, Guangdong, China
202220112193@mail.scut.edu.cn

Zhiqi Lin
South China University of Technology
School of computer Science & Engineering
Guangzhou, Guangdong, China
202311089192@mail.scut.edu.cn

Delu Zeng*
South China University of Technology
School of Electronic and Information Engineering
Guangzhou, Guangdong, China
dlzeng@scut.edu.cn

Abstract

In the field of blind image quality assessment, accurately capturing localized distortions and structural inconsistencies within images remains a significant challenge. To tackle this issue, we propose BlockIQA, a novel framework that enhances local sensitivity through deep block analysis. BlockIQA divides images into non-overlapping blocks and employs a multi-branch architecture that integrates ResNet50, the Feature Pyramid Network, and an Auxiliary Feature Extraction Layer. The primary innovations of this paper include: (1) Segmenting images into smaller blocks for detailed analysis and employing a Gaussian similarity model that dynamically adapts to variations in feature dimensions and directional consistency. This approach enables more precise characterization of local image features. (2) Achieving a balance between global semantic information and localized distortion patterns through multiscale feature fusion using feature pyramid networks and auxiliary feature extraction layer. This ensures that while capturing overall image semantics, no fine local distortion information is overlooked. Experimental results demonstrate that BlockIQA performs well across datasets with various types of distortions and exhibits strong generalizability across different databases. In summary, BlockIQA pioneers a new deep learning architectural paradigm for blind image quality assessment. Its design philosophy of enhancing local sensitivity through deep block analysis provides valuable new ideas and methods for research and practice in this domain.

CCS Concepts

• **Computing methodologies → Image processing**.

Keywords

Image quality assessment; no reference/blind; image block analysis; multiscale feature fusion; local sensitivity enhancement; feature pyramid network; gaussian similarity

ACM Reference Format:
Yuqi Pang, Yican Liu, Zhiqi Lin, and Delu Zeng. 2025. BlockIQA: Local Sensitivity-Enhanced Blind Image Quality Assessment through Deep Block Analysis. In *Proceedings of the 2025 International Conference on Multimedia Retrieval (ICMR '25), June 30-July 3, 2025, Chicago, IL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3731715.3733280

1 Introduction

Image quality assessment (IQA) techniques aim to simulate the human eye's perception of image quality. Depending on the degree of dependence on the reference image, image quality assessment methods are broadly categorized into three paradigms: full-reference (FR-IQA), reduced-reference (RR-IQA), and no-reference (NR-IQA). The FR-IQA method requires access to the original undistorted image for comparison. Early methods such as Structural Similarity Index (SSIM) [36] measure the similarity between the reference and distorted images in terms of brightness, contrast and structure. More advanced methods such as Visual Information Fidelity (VIF) [29] quantify the loss of information in distorted images. FR-IQA, meanwhile, achieves high accuracy (for example, WaDIQaM [24] achieves SRCC=0.955 on CSIQ), but its dependence on the reference image limits real-world applicability [35], [37]. RR-IQA methods utilize partial reference information (for example, statistical features) to assess quality. The reduced reference difference (RRED) [33] utilizes the entropy difference between the reference image and the distorted image. However, RRIQA still requires partial reference data, which makes it unsuitable for scenarios where the reference image is completely unavailable. NR-IQA, on the other hand, does not require a reference image and can assess the quality based only on the own features of the distorted image, making it the preferred choice for practical applications, since we do not have access to a reference image in the real world [22].

In the digital era, visual content quality profoundly impacts user experience across applications like video streaming and telemedicine. NR-IQA aims to predict perceptual quality without reference images, yet accurately capturing localized distortions (e.g., regional

[*] Corresponding author.

ICMR '25, Chicago, IL, USA

ACM ISBN 979-8-4007-1877-9/2025/06
https://doi.org/10.1145/3731715.3733280

blur or sensor noise) remains challenging. While deep learning has advanced NR-IQA through CNNs [13], [14], [19] and transformers [23], [21], most existing deep learning-based NR-IQA methods, such as HyperIQA [26], rely on global feature pools, which obscures local anomalies [34]. This limitation becomes critical for real-world images with non-uniform distortions, where regional quality varies significantly [3].

To address this, we propose BlockIQA, a framework enhancing local sensitivity via deep block analysis. Unlike MANIQA [27], which uses rigid patch selection, our dynamic blocking adapts to image resolution and content. Furthermore, we replace conventional cosine similarity with a Gaussian metric that jointly models feature magnitude and direction, overcoming scale sensitivity issues [7], [28]. Experiments on both synthetic (TID2013) and authentic (CLIVE) datasets validate BlockIQA's superiority, achieving state-of-the-art SRCC scores of 0.985 on CSIQ. Fig. 1 illustrates the local sensitivity of our method and the experimental results compared to other methods in the TID2013 dataset.

The main contributions of this paper can be summarized as follows:

- We propose the BlockIQA architecture, an innovative deep neural network for processing image blocks to comprehensively assess image quality, achieving local sensitivity-enhanced blind image quality assessment via deep block analysis.
- We utilize pre-trained ResNet50 models and Feature Pyramid Networks(FPN) to extract multiscale features from image blocks, enhancing the ability to capture local details and global information. In addition, an auxiliary module is introduced to further improve assessment performance.
- We conduct extensive dataset experiments to confirm BlockIQA's effectiveness, reliability, and practicality with its focus on local sensitivity enhancement.

2 Related Work

NR-IQA is a critical research area in image processing and computer vision, aiming to predict perceptual quality without reference images. This capability is essential for applications such as video streaming, telemedicine, and image enhancement, where reference images are often unavailable. NR-IQA methods can be broadly categorized into two paradigms: supervised and unsupervised learning. Supervised methods rely on subjective quality scores for training, while unsupervised methods operate without such annotations. Recently, NR-IQA can be grouped into three main approaches: (1) Handcrafted feature-based methods, such as BRISQUE [1], which leverage natural scene statistics (NSS) but lack semantic awareness; (2) CNN-based methods, including DBCNN [32] and MetaIQA [10], which extract hierarchical features but often struggle with localized distortions; and (3) Transformer-based frameworks, such as MUSIQ [12], which excel in capturing long-range dependencies but are computationally intensive for fine-grained analysis.

Traditional machine learning approaches, such as NSS-based methods [17], [30], assume that pristine images follow specific statistical distributions, which are disrupted by distortions. While effective, these methods often fail to capture high-level semantic information. Deep learning-based approaches, such as MUSIQ [12] and AHIQ [25], address this limitation by leveraging pre-trained CNN models to extract rich features for quality assessment. However, most existing methods rely on global feature pooling, which can obscure local distortions, especially in real-world images with non-uniform quality degradation.

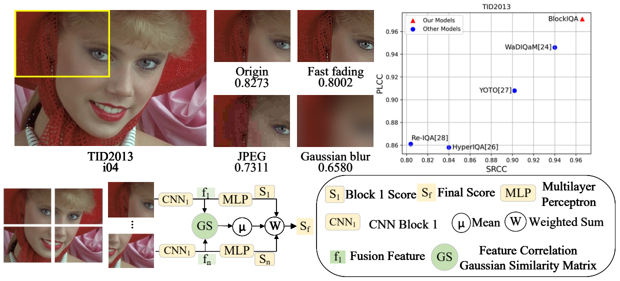


Figure 1: This figure illustrates the innovativeness of our approach. There are obvious differences in the scores of local images with different distortion processes, which proves that this method can enhance local sensitivity. In addition, the results on TID2013 dataset also demonstrate the effectiveness and accuracy of our methodology.

Deep learning algorithms focus on designing different deep neural network architectures, such as using deep convolutional neural networks to extract deep features of images, providing rich information for the assessment of image quality [38], [18]. Notably, Align-IQA [40] introduced customizable guidance for human preference alignment but used fixed block sizes, limiting adaptability. Meanwhile, YOTO [41] unified FR/NR-IQA paradigms but neglected interblock relationships. Our work bridges these gaps by integrating dynamic blocking with Gaussian similarity, enabling precise localization of distortions while maintaining global consistency.

The success of deep learning in NR-IQA is largely attributed to architectures like ResNet [11] and FPN [16]. ResNet, with its residual connections, addresses the vanishing gradient problem, enabling the training of deeper networks. FPN, on the other hand, enhances multiscale feature extraction by combining low-resolution, semantically rich features with high-resolution, detail-rich features through a top-down pathway and lateral connections. These advancements have significantly improved the performance of tasks such as object detection and image quality assessment.

In summary, while existing NR-IQA methods have made significant progress, they often fail to balance local distortion sensitivity and global semantic understanding. Our proposed BlockIQA framework addresses this limitation by integrating dynamic blocking, multiscale feature fusion, and Gaussian similarity, offering a robust solution for blind image quality assessment.

3 Proposed Method

The BlockIQA framework aims to achieve precise blind image quality assessment by deeply analyzing the Gaussian similarity between image blocks, with a focus on local sensitivity. The framework consists of three key stages: image block partitioning, multiscale feature extraction and fusion, and quality score regression. First, the input image is divided into local blocks to capture fine-grained distortions. Then, these blocks are processed by a multiscale feature extraction module, which integrates a pre-trained ResNet50 backbone, a FPN, and an Auxiliary Feature Extraction Layer (AFEL). The extracted features are adaptively fused to compute a Gaussian similarity matrix, modeling the structural relationships between blocks. Finally, a regression network aggregates these features to predict the overall image quality score. This hierarchical design ensures both local distortion sensitivity and global structural consistency. The overall technical workflow of the framework is shown in Fig. 2. In the following sections, we will detail each component within the BlockIQA framework.

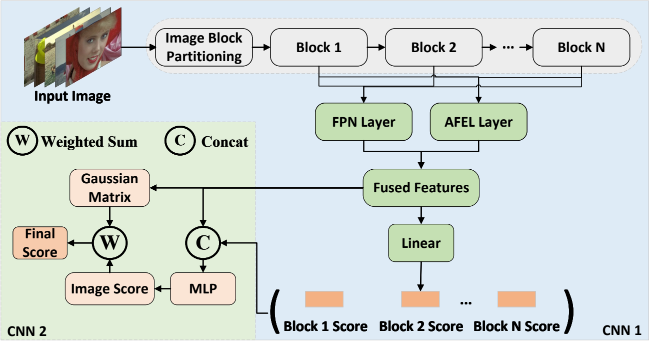


Figure 2: The diagram illustrates the overall process framework of BlockIQA, showing how distorted images can be processed through components such as image partitioning, CNN1, and CNN2 to generate fusion features, and to obtain block scores and final image quality scores. The FPN layer and the AFEL layer are described in detail in Fig. 4 and 5.

3.1 Partitioning of Image Blocks and Gaussian Similarity Selection

To intuitively demonstrate the necessity of image blocking, we visualized the distribution of interblock similarity in images of varying quality using heatmaps (Fig. 3). In the heatmaps, brighter colors indicate higher interblock similarity, while darker colors signify lower similarity. The experiment selected three typical samples:the thermograms of high-quality images (score 6.31) showed an overall uniformly high brightness (Fig. 3, left), indicating a high degree of interblock structural consistency without significant local distortion; the thermograms of medium-quality images (score 4) showed a significant decrease in the brightness of local regions (Fig. 3, middle), reflecting a decrease in interblock similarity, which corresponds to a local distortion such as blurring or noise; and the thermograms of low-quality images (score 2.5) have overall dark and alternately bright and dark heatmaps (Fig. 3, right), showing a chaotic global structure with drastic differences in interblock similarity. By comparing these heatmaps, it is evident that interblock similarity is closely related to image quality. The blocking strategy can effectively locate local distortions, avoiding the smoothing effect of global features on local anomalies, thereby providing a more accurate basis for image quality assessment.

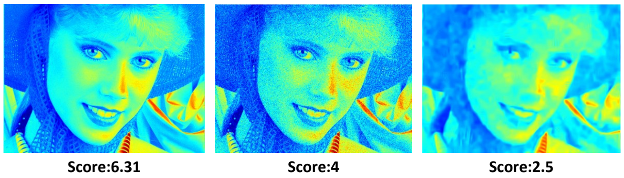


Figure 3: Heatmaps of three images with different distortion levels selected from TID2013 are shown, with scores decreasing from left to right. In the heatmaps, brighter colors indicate higher interblock similarity, while darker colors indicate lower similarity.

In addition, selecting an appropriate similarity measure is also crucial. We compared Gaussian similarity and cosine similarity and found that Gaussian similarity outperforms cosine similarity both theoretically and empirically. Cosine similarity only measures the directional consistency of feature vectors and is defined as:

$$\cos(f_i, f_j) = \frac{f_i \cdot f_j}{\|f_i\|\|f_j\|} \tag{1}$$

However, it ignores the differences in feature magnitudes and is sensitive to scale variations. For example, features with similar directions but significantly different magnitudes (such as strong noise versus weak noise) would still be judged as highly similar. This limits its applicability in image quality assessment. In contrast, Gaussian similarity integrates both directional and Euclidean distance information and is defined as:

$$GS(i,j) = \exp\left(-\frac{\|f_i - f_j\|^2}{2\sigma^2}\right) \tag{2}$$

The bandwidth parameter σ is dynamically adjusted as:

$$\sigma = \sqrt{\frac{\dim(f_i) + \dim(f_j)}{2} \cdot (\frac{1 + \cos(f_i, f_j)}{2} + k)} \tag{3}$$

where $\dim(\cdot)$ denotes the dimension of the features, f_i and f_j denote the features extracted from the i-th and j-th image block, respectively, and k is a positive number infinitely close to zero.

This design allows the Gaussian similarity to increase σ when the feature directions are consistent ($\cos \approx 1$), thereby reducing the impact of magnitude differences. Conversely, it decreases σ when the directions are inconsistent ($\cos \approx -1$), thereby amplifying the distance differences. Moreover, through normalization by $\dim(f)$, Gaussian similarity effectively mitigates the issue of Euclidean distance inflation in high-dimensional spaces, thus demonstrating stronger robustness in high-dimensional feature spaces. From a theoretical perspective, the Gaussian kernel function can be regarded as a local neighborhood measure in high-dimensional feature spaces, which aligns more closely with the non-linear manifold assumption of visual features. This viewpoint is based on manifold learning theory, which emphasizes that visual features typically reside on low-dimensional manifolds within high-dimensional spaces. Gaussian similarity is thus better suited to capture such local structures.

In the image blocking strategy, the input image is divided into $n \times n$ non-overlapping blocks, each measuring $\frac{H}{n} \times \frac{W}{n}$. This approach balances local detail capture and contextual information retention. Each block is processed independently to avoid cross-block interference, enhancing local feature analysis. For example, local blurring or noise affects specific blocks, causing noticeable deviations in their features. By analyzing each block separately, the model avoids detail loss from global pooling, providing a more nuanced assessment.

Heatmap analysis reveals a strong correlation between local distortions and interblock similarity. The strategy integrates content-adaptive blocking and Gaussian similarity, significantly enhancing sensitivity to local distortions. Gaussian similarity's dynamic bandwidth and dimension normalization make it superior to traditional cosine similarity in complex scenarios, a core innovation of BlockIQA. Future work may involve dynamically adjusting block size based on image resolution and content complexity, with smaller blocks for high-resolution images and larger blocks for low-resolution ones. This could be combined with semantic segmentation for further improvements.

3.2 multiscale Feature Extraction and Fusion

To better extract multiscale features from image blocks, we designed a FPN that enables the model to capture both local details and global context, as shown in Fig. 4. This process begins with the ResNet50 backbone, which generates feature maps at different scales. These feature maps correspond to different network depths and resolutions. Specifically, feature maps from shallower layers have higher resolutions and shallower semantic information, while those from deeper layers have lower resolutions and deeper semantic information. For example, in ResNet50, four stages (from stage 2 to stage 5) of feature maps C_1, C_2, C_3, and C_4 are typically generated, with resolutions of $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$, and $\frac{1}{32}$ of the input image, respectively. Subsequently, the FPN constructs the feature pyramid through lateral connections and a top-down pathway. Lateral connections fuse feature maps from the same level, while the top-down pathway restores the resolution of feature maps through upsampling.

As shown in Fig. 4, for each feature map layer, the FPN applies a convolutional operation to generate an intermediate feature map. For example, for the feature map C_i at layer i, a 1×1 convolution is applied to reduce the number of channels, resulting in an intermediate feature map P_i. This process can be represented as:

$$P_i = \mathrm{conv}_{1\times1}(C_i) \tag{4}$$

Subsequently, the resolution of the intermediate feature map is enhanced through upsampling. Specifically, the feature map P_{i+1} from the higher layer is upsampled to match the resolution of C_i, and then added element-wise to P_i to obtain the fused feature map. This process can be expressed as follows:

$$P_i' = \mathrm{upsample}(P_{i+1}) + P_i \tag{5}$$

In the formula, "upsample" represents the upsampling operation, which is typically bilinear interpolation or nearest neighbor interpolation. The element-wise addition operation combines the low-resolution feature map with the high-resolution feature map, thereby enhancing the spatial details and semantic information of the feature map.

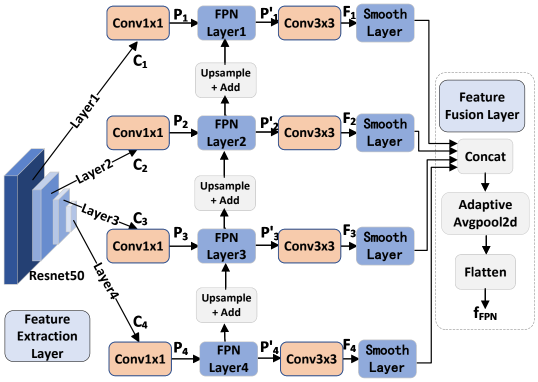


Figure 4: FPN layer. It shows the process of extracting the features of image patches from different layers of the pre-trained ResNet50 and fusing them to obtain multiscale features. The feature C_i of each layer undergoes a 1×1 convolution to obtain P_i, and P_i is combined with the upsampled P_{i+1} from the previous layer. Subsequently, a 3×3 convolution is performed to obtain F_i. Finally, all the features are combined to obtain f_{FPN}.

Subsequently, the fused feature map is convolved again to generate the final feature pyramid layer. For example, for each fused feature map P_i, a 3×3 convolution is applied to smooth the feature map and generate the final feature map F_i. This process can be represented as:

$$F_i = \mathrm{conv}_{3\times3}(P_i') \tag{6}$$

Ultimately, the feature extracted from the FPN layer for each image block is represented as:

$$f_{FPN} = \mathrm{concat}(F_1, F_2, F_3, F_4) \tag{7}$$

Here, F_1, F_2, F_3, and F_4 correspond to the feature maps generated from different levels of the FPN, capturing multiscale information. The concatenation operation integrates these feature maps into a comprehensive representation f_{FPN}, which encapsulates both local details and global context for each image block.

In summary, the Feature Pyramid Network builds a robust feature pyramid by integrating feature maps from different levels for multi-scale feature extraction. This approach has played a significant role in improving the model's ability to detect objects at different scales and has achieved notable results in tasks such as image quality assessment.

3.3 Auxiliary feature extraction

In the BlockIQA framework, the AFEL is a lightweight convolutional module that is utilised to enhance the sensitivity of the model to local distortions. As shown in Fig. 5, AFEL captures the underlying details of the image through shallow feature extraction, thus complementing the FPN in multiscale global feature extraction.

The core design of the AFEL layer includes convolutional operations, nonlinear activation, adaptive pooling, and feature transformation. Firstly, feature extraction is performed by a 3×3 convolutional kernel to extract local features (e.g., edges, texture, and noise patterns) of the image, and fast downsampling is performed using a large step size to reduce the computational complexity. Subsequently, the ReLU activation function introduces nonlinearities to enhance the model's ability to model complex distortion patterns. Adaptive mean pooling further aggregates spatial information, removes location sensitivity, and preserves statistical features in the channel dimension. Finally, the features are converted into high-dimensional semantic representations through the fully connected layer to capture the abstract patterns of local distortion and further enhance the flexibility of feature representation through the ReLU activation function.

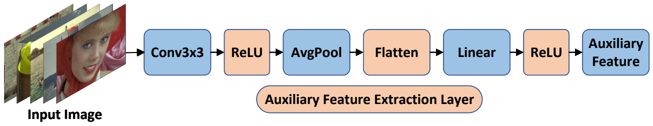


Figure 5: AFEL layer. It is used to extract shallow features to complement FPN in multiscale global feature extraction.

We denote the feature extracted from the AFEL as f_{AFEL}, consequently, the total feature for each image block is represented as:

$$f_i = f_{\text{FPN}} + f_{\text{AFEL}} \tag{8}$$

AFEL is significant in the BlockIQA framework. It significantly enhances the model's ability to perceive local distortions and makes up for the inadequacy of FPN for small-area local distortions (e.g., block artefacts and subtle noise). Experiments show that the addition of AFEL improves the SRCC of the model by 2.3% on the TID2013 dataset. In addition, the joint features formed by fusing AFEL features with FPN features achieve more comprehensive quality assessment using global semantic information and local statistics. Ablation experiments show that removing AFEL leads to a 1.8% decrease in PLCC on the CLIVE dataset. In addition, the design of the AFEL layer optimises computational efficiency by including only lightweight operations (single-layer convolution + full connectivity) with a parameter count of $64 \times 3 \times 3 \times 3 + 64 \times 128 = 2.3\text{K}$, which accounts for 0.6% of the total model parameters and adds almost no computational overhead.

In summary, AFEL significantly improves BlockIQA's sensitivity to complex distortions through efficient local feature extraction and fusion mechanisms, balances computational efficiency and feature expression capability, and provides a new paradigm of local-global synergistic analysis for reference-free image quality assessment.

3.4 Linear Regression

In the BlockIQA architecture, we cleverly integrate multiscale features through a fully connected layer to generate the final quality score of the image. This step receives features processed by CNN Block2 and converts them into a continuous predicted score through a linear layer, representing the estimated quality of the image. If we define f_i as a feature extracted from the i-th image block, G as the aggregator function, and use the sigmoid activation function, then the fraction S_i of the i-th image block can be expressed as:

$$S_i = \text{Sigmoid}(G(f_i)) \tag{9}$$

In this process, we employ a Multi-Layer Perceptron (MLP) as the aggregator to deeply process the fused features. The aggregator consists of components such as linear layers, batch normalization, ReLU activation functions, and Dropout. At the same time, after obtaining the Gaussian similarity matrix, we will calculate the final features F of each image block as follows:

$$\text{Mean}(F_i) = \frac{1}{n}\sum_{j=1}^{n} GS(i,j) \tag{10}$$

The image block score reflects the local quality assessment, and the features contain the multi-scale information of the image, which can integrate complementary information and achieve a more comprehensive quality assessment. Therefore, our paper uses a weighted sum to obtain the final image quality score:

$$S_f = \frac{1}{n}\sum_{i=1}^{n} S_i \cdot \text{Mean}(F_i) \tag{11}$$

To ensure the accuracy of model predictions, we employ the Mean Squared Error (MSE) loss function to measure the deviation between the model's predicted scores and the ground quality scores. The MSE is defined by the following formula:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n} (S_i - \hat{S}_i)^2 \tag{12}$$

Here, n represents the total number of samples, S_i is the actual quality score of the i-th sample, and $\hat{S}_i$ is the predicted score of the same sample by the model. Through algorithms such as backpropagation and gradient descent, the model continuously iterates until it achieves the best performance on the validation set. This process enables BlockIQA to automatically learn image features and accurately predict image quality, providing an efficient and precise method for image quality assessment. It not only improves the accuracy of the scores but also enhances the efficiency of the evaluation process.

4 Experiments

4.1 Datasets and Evaluation Protocols

Datasets. To address the limitations of earlier databases that often contained only one type of distortion, this researchers have turned to "in-the-wild" datasets such as CLIVE[5], CID2013[31], and SPAQ[39], which include more distortions from the real world. Among them, CLIVE contains 1,162 real distortion images captured by various mobile devices. CID2013 consists of 480 images taken by 79 imaging devices, used to study commercially relevant distortion issues. SPAQ includes 11,000 images captured by 66 mobile devices, each annotated with brightness, content tags, and EXIF data. Additionally, we evaluated our method on four traditional synthetic distortion datasets: LIVE[9], TID2013[20], CSIQ[15], and KADID[8]. The LIVE dataset includes 779 images with five types of

Table 1: Comparison results of our approach with several state-of-the-art NR-IQA methods on different datasets. The best result is marked in red, and the second best result is marked in blue.

Dataset	Synthesis								Authentic					
	CSIQ		LIVE		TID2013		KADID		CLIVE		CID2013		SPAQ	
Criterion	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
BRISQUE[1]	0.746	0.829	0.939	0.935	0.604	0.694	0.528	0.567	0.608	0.629	0.615	0.648	0.809	0.817
FRIQUEE[6]	0.835	0.874	0.940	0.944	0.680	0.753	-	-	0.682	0.705	-	-	-	-
DBCNN[32]	0.946	0.959	0.968	0.971	0.816	0.865	0.878	0.878	0.851	0.869	-	-	0.911	0.915
WaDIQaM[24]	0.955	0.973	0.954	0.963	0.940	0.946	0.752	0.739	0.671	0.680	0.708	0.729	0.840	0.845
MetaIQA[10]	0.899	0.908	0.960	0.959	0.856	0.868	0.762	0.775	0.835	0.802	0.766	0.784	-	-
HyperIQA[26]	0.923	0.942	0.962	0.968	0.840	0.858	0.852	0.845	0.859	0.882	0.490	0.612	0.916	0.919
YOTO[41]	0.952	0.962	0.974	0.976	0.902	0.908	0.885	0.884	0.841	0.892	-	-	-	-
Re-IQA[2]	0.947	0.960	0.970	0.971	0.804	0.861	0.872	0.885	0.840	0.854	-	-	0.900	0.918
SADCIQA[4]	0.957	0.965	0.969	0.978	0.856	0.882	0.901	0.914	0.850	0.857	-	-	0.916	0.920
Align-IQA[40]	0.975	0.981	0.985	0.987	0.955	0.960	0.928	0.932	0.905	0.916	-	-	-	-
AHIQ[25]	0.984	0.989	0.975	0.978	0.962	0.968	-	-	-	-	-	-	-	-
MANIQA[27]	0.961	0.968	0.982	0.983	0.937	0.943	0.944	0.946	-	-	-	-	-	-
BlockIQA	0.985	0.988	0.981	0.982	0.966	0.971	0.970	0.972	0.875	0.892	0.941	0.938	0.917	0.920

distortions. TID2013 contains 3,000 images with 24 different types of distortions. The CSIQ dataset includes 866 images with six types of distortions. The KADID dataset consists of 10,125 images using 25 different distortion techniques.

Evaluation Metrics. To assess the model's performance across various IQA databases, we employed the Spearman Rank Correlation Coefficient (SRCC) and the Pearson Linear Correlation Coefficient (PLCC) as evaluation metrics. Both measures range from -1 to 1, with higher values indicating better performance. SRCC measures the monotonic relationship between true values and predicted scores, while PLCC quantifies the degree of linear correlation between them.

Implementation Details. During training, we utilized the mean squared error loss function and the Adam optimizer, with an initial learning rate set to 0.0001 and weight decay to prevent overfitting. The datasets were split into 80% for training and 20% for testing. We implemented an exponential learning rate decay strategy, multiplying the learning rate by a decay factor after a certain number of epochs to dynamically adjust the learning rate. Additionally, to ensure the comparability of the model's evaluation results, we normalized the scores of each image in the datasets to a range of 0 to 1.

4.2 Experiment Results

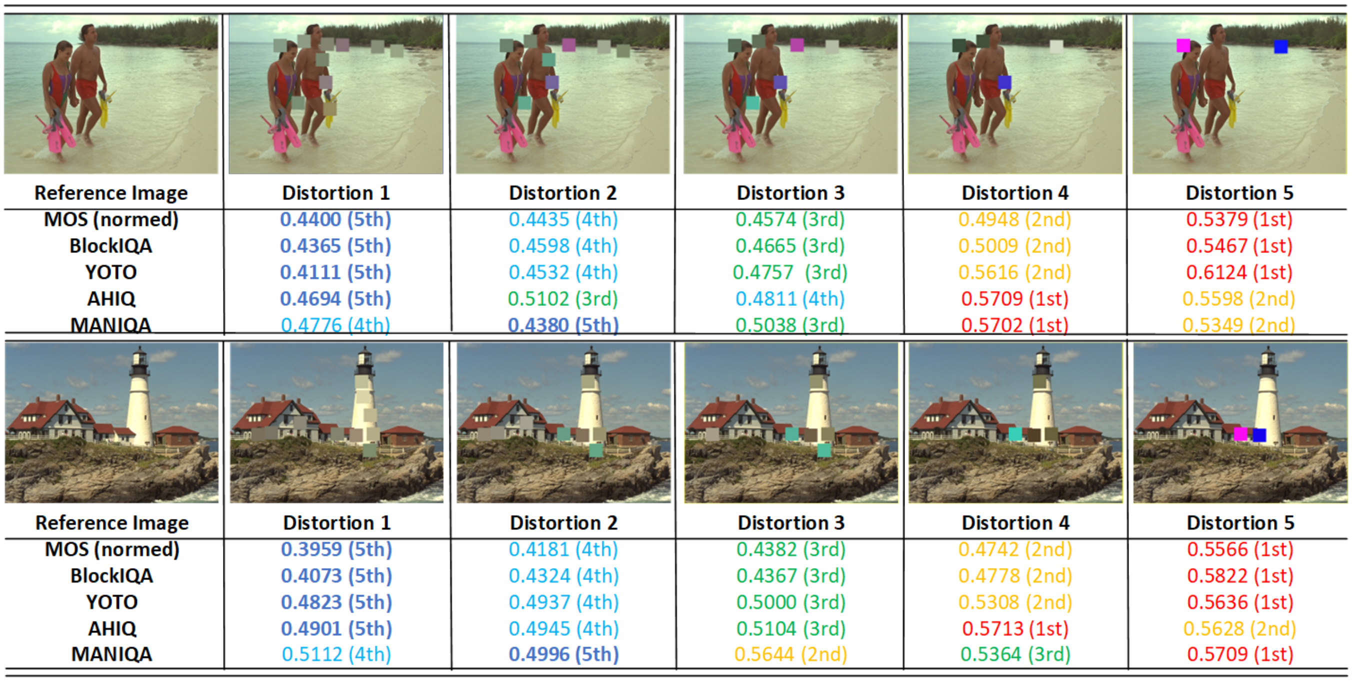
We compared BlockIQA against state-of-the-art NR-IQA methods on both synthetic and authentic distortion datasets. As shown in Tab. 1, BlockIQA consistently outperforms competing models across all evaluation metrics. For example, on the CSIQ dataset, BlockIQA achieves an SRCC of 0.985, surpassing the second-best model (Align-IQA) by 1.0%. Similarly, on the LIVE dataset, BlockIQA achieves an SRCC of 0.981, demonstrating its superior ability to handle synthetic distortions.

On authentic distortion datasets such as CLIVE and CID2013, BlockIQA also exhibits strong performance, with SRCC scores of 0.875 and 0.941, respectively. This can be attributed to the model's ability to independently analyze regions with varying distortion levels, as illustrated in Fig. 6. For instance, BlockIQA's ranking of images with the same type of distortion is highly consistent with human evaluations, even under extreme distortion conditions. This robustness is particularly evident in comparisons with models like YOTO and MANIQA, where BlockIQA's scores align more closely with subjective assessments.

4.3 Comparative Assessment

Different Number of Image Blocks. Tab. 2 presents the performance comparison with different block configurations (n = 2, 3, 4). As shown, increasing the number of blocks does not always lead to improved performance. For instance, on the CSIQ dataset, the SRCC and PLCC values decreased from 0.985 and 0.988 (n=2) to 0.982 and 0.986 (n=3), and further to 0.979 and 0.982 (n=4). Similar trends were observed on other datasets such as LIVE, TID2013, and KADID. This suggests that while dividing images into more blocks can enhance local sensitivity, excessive partitioning may weaken the model's ability to capture global semantic information, leading to a decline in overall performance. Additionally, increasing the number of blocks also increases computational complexity, potentially extending training and inference times. Therefore, it is crucial to strike a balance between local sensitivity and computational efficiency when determining the number of image blocks.

Gaussian Similarity vs. Cosine Similarity. Tab. 3 presents the performance comparison between Gaussian similarity and cosine similarity on the same dataset. As shown in the table, Gaussian similarity consistently outperforms cosine similarity across all datasets in terms of both SRCC and PLCC. For example, on the CSIQ dataset, Gaussian similarity achieves an SRCC of 0.985 and a PLCC of 0.988, while cosine similarity only reaches an SRCC of 0.978 and a PLCC of 0.987. Similar trends are observed on the LIVE, CLIVE, and SPAQ datasets. This indicates that Gaussian similarity is more effective



Reference Image	Distortion 1	Distortion 2	Distortion 3	Distortion 4	Distortion 5
MOS (normed)	0.4400 (5th)	0.4435 (4th)	0.4574 (3rd)	0.4948 (2nd)	0.5379 (1st)
BlockIQA	0.4365 (5th)	0.4598 (4th)	0.4665 (3rd)	0.5009 (2nd)	0.5467 (1st)
YOTO	0.4111 (5th)	0.4532 (4th)	0.4757 (3rd)	0.5616 (2nd)	0.6124 (1st)
AHIQ	0.4694 (5th)	0.5102 (3rd)	0.4811 (4th)	0.5709 (1st)	0.5598 (2nd)
MANIQA	0.4776 (4th)	0.4380 (5th)	0.5038 (3rd)	0.5702 (1st)	0.5349 (2nd)

Reference Image	Distortion 1	Distortion 2	Distortion 3	Distortion 4	Distortion 5
MOS (normed)	0.3959 (5th)	0.4181 (4th)	0.4382 (3rd)	0.4742 (2nd)	0.5566 (1st)
BlockIQA	0.4073 (5th)	0.4324 (4th)	0.4367 (3rd)	0.4778 (2nd)	0.5822 (1st)
YOTO	0.4823 (5th)	0.4937 (4th)	0.5000 (3rd)	0.5308 (2nd)	0.5636 (1st)
AHIQ	0.4901 (5th)	0.4945 (4th)	0.5104 (3rd)	0.5713 (1st)	0.5628 (2nd)
MANIQA	0.5112 (4th)	0.4996 (5th)	0.5644 (2nd)	0.5364 (3rd)	0.5709 (1st)

Figure 6: Quality score comparison for BlockIQA against other state-of-the-art models YOTO, AHIQ and MANIQA. From left to right: the original image and 5 distorted images with decreasing distortion levels. Different color codes are applied for better visualization of the 1st, 2nd, 3rd, 4th, and 5th ranking of prediction scores.

Table 2: Performance of different number of image blocks on the same dataset. The Best Result is Presented with Boldfase.

Dataset	n=2		n=3		n=4	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
CSIQ	**0.985**	**0.988**	0.982	0.986	0.979	0.982
LIVE	**0.981**	**0.982**	0.975	0.976	0.972	0.975
TID2013	**0.966**	**0.971**	0.942	0.954	0.949	0.958
KADID	**0.970**	**0.972**	0.961	0.962	0.950	0.953
CLIVE	**0.875**	**0.892**	0.832	0.848	0.805	0.832
CID2013	**0.941**	**0.938**	0.891	0.903	0.898	0.901
SPAQ	**0.917**	**0.920**	0.907	0.913	0.908	0.915

Table 3: Performance with different similarities on the same dataset. The Best Result is Presented with Boldfase.

Dataset	Cosine Similarity		Gaussian Similarity	
	SRCC	PLCC	SRCC	PLCC
CSIQ	0.978	0.987	**0.985**	**0.988**
LIVE	0.972	0.973	**0.981**	**0.982**
CLIVE	0.814	0.849	**0.875**	**0.892**
SPAQ	0.904	0.908	**0.917**	**0.920**

in capturing the local structural information and has a stronger correlation with human perception of image quality. The results demonstrate the superiority of Gaussian similarity over cosine similarity in the context of image quality assessment.

Dynamic σ vs. Fixed σ. Tab. 4 presents the performance comparison between dynamic σ and fixed σ on the same dataset. The square of our fixed σ is set to the dimension of the feature, while the dynamic σ changes according to the different similarities between the features. As shown in the table, dynamic σ consistently outperforms fixed σ across all datasets in terms of both SRCC and PLCC. This indicates that it can adjust its size according to the consistency of the direction of the eigenvector, increase σ to reduce the influence of amplitude difference when the feature direction is consistent, and reduce σ to amplify the distance difference when the direction is inconsistent, so as to accurately identify the local distortion.

Table 4: Performance with different σ on the same dataset. The Best Result is Presented with Boldfase.

Dataset	Fixed σ		Dynamic σ	
	SRCC	PLCC	SRCC	PLCC
CSIQ	0.976	0.978	**0.985**	**0.988**
LIVE	0.974	0.976	**0.981**	**0.982**
CLIVE	0.868	0.883	**0.875**	**0.892**
SPAQ	0.905	0.910	**0.917**	**0.920**

4.4 Ablation Experiments

As shown in Tab. 5, we conducted separate experiments to analyze the effectiveness of each component of the proposed BlockIQA. Our baseline model uses only FPN as the feature extraction framework. The results show that our image partitioning module (BP) effectively enhances the model's ability to capture local distortions by dividing images into blocks, avoiding the smoothing effect of global features on local anomalies. Our AFEL module effectively improves the model's sensitivity to local distortions by extracting underlying details of the image through shallow feature extraction, compensating for the limitations of FPN in feature extraction. By combining both the image partitioning module and the auxiliary feature extraction module, the model achieves a more comprehensive quality assessment, utilizing both global semantic information and local statistics. This combination further improves the accuracy of image quality assessment, demonstrating the effectiveness of each component in enhancing the overall performance of BlockIQA.

Table 5: Ablation study results. The Best Result is Presented with Boldfase.

BP	AFEL	TID2013		CLIVE	
		SRCC	PLCC	SRCC	PLCC
✗	✗	0.950	0.958	0.852	0.865
✓	✗	0.957	0.966	0.864	0.884
✗	✓	0.954	0.963	0.854	0.870
✓	✓	**0.966**	**0.971**	**0.875**	**0.892**

4.5 Cross Dataset Evaluations

To test the generalization ability of our model across different databases, we conducted cross-dataset evaluations. The results are presented in Tab. 6. As shown, BlockIQA outperformed other models on all synthetic distortion datasets and also demonstrated excellent performance on other datasets. For example, when trained on the LIVE dataset and tested on CSIQ, BlockIQA scored 4.1% higher than the second-best model. When trained on the CSIQ dataset and tested on LIVE, BlockIQA scored 5.5% higher than the second-best model. Similarly, when trained on the CLIVE dataset and tested on TID2013, BlockIQA scored 3% higher than the second-best model. These results indicate that our model shows high SRCC values across various dataset combinations, particularly in tests between synthetic datasets, demonstrating the model's robust generalization capability and excellent cross-database assessment performance.

Table 6: SRCC Results in A Cross-Database Setting. The Best Result is Presented with Boldfase.

Training	Testing	BRISQUE	FRIQUEE	DBCNN	Ours
CSIQ	LIVE	0.847	0.879	0.877	**0.934**
LIVE	CSIQ	0.562	0.722	0.758	**0.799**
TID2013	CLIVE	0.254	0.181	**0.457**	0.449
CLIVE	TID2013	0.280	0.424	0.424	**0.454**

5 Conclusion

In conclusion, this paper introduces BlockIQA, a novel reference-free image quality assessment framework for local sensitivity-enhanced blind image quality assessment via deep block analysis. It combines image block analysis and multiscale feature fusion. BlockIQA utilizes pre-trained ResNet50, FPN and AFEL to ensure robust feature extraction at multiple scales while capturing local details and global semantics. It also employs a Gaussian similarity metric dynamically adjusted to accommodate feature dimensionality and orientation consistency to accurately assess image quality. Experimental results confirm its excellent performance on a variety of datasets as well as its generalization ability in cross-database evaluation. Future work focuses on optimizing the number of image blocks to achieve a balance between performance and computational efficiency. In addition, we plan to utilize the method of object detection to achieve the adaptability of image block segmentation, as well as assign weights to different blocks according to their importance. BlockIQA provides a new accurate and efficient tool for image quality assessment, and we look forward to its further application and development.

Acknowledgments

This work was supported in part by the Fundamental Research Program of Guangdong, China, under Grant 2020B1515310023 and Grant 2023A1515011281.

References

[1] A. K. Moorthy A. Mittal and A. C. Bovik. 2012. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing* 21, 12 (2012), 4695–4708.

[2] S. Mishra A. Saha and A. C. Bovik. 2023. Re-iqa: Unsupervised learning for image quality assessment in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 5846–5855.

[3] A. C. Bovik. 2013. Automatic prediction of perceptual image and video quality. *Proc. IEEE* 101, 9 (2013), 2008–2024.

[4] Menglong Chen, Jianming Wang, Zhitao Xiao, and Yukuan Sun. 2024. A Saliency-Aware NR-IQA Method by Fusing Distortion Class Information. In *International Conference on Pattern Recognition.* Springer, 130–143.

[5] D. Ghadiyaram and A. C. Bovik. 2015. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing* 25, 1 (2015), 372–387.

[6] D. Ghadiyaram and A. C. Bovik. 2017. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of vision* 17, 1 (2017), 32–32.

[7] Paolo Giannitrapani, Elio D Di Claudio, and Giovanni Jacovitti. 2025. Full-reference calibration-free image quality assessment. *Signal Processing: Image Communication* 130 (2025), 117212.

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

[9] M. F. Sabir H. R. Sheikh and A. C. Bovik. 2006. A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on image processing* 15, 11 (2006), 3440–3451.

[10] L. Li H. Zhu, W. Dong J. Wu, and G. Shi. 2020. MetaIQA: Deep meta-learning for no-reference image quality assessment. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition.* 14143–14152.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 770–778.

[12] Q. Wang J. Ke, P. Milanfar Y. Wang, and F. Yang. 2021. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision.* 5148–5157.

[13] Le Kang, Peng Ye, Yi Li, and David Doermann. 2014. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition.* 1733–1740.

[14] Jongyoo Kim and Sanghoon Lee. 2016. Fully deep blind image quality predictor. *IEEE Journal of selected topics in signal processing* 11, 1 (2016), 206–220.

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

[16] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. 2017. Feature pyramid networks for object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2117–2125.

[17] A. C. Bovik M. A. Saad and C. Charrier. 2012. Blind image quality assessment: A natural scene statistics approach in the DCT domain. *IEEE transactions on Image Processing* 21, 8 (2012), 3339–3352.

[18] C. Cai M. H. Eybposh, J. Rodriguez-Romaguera A. Moossavi, and N. C. Pegard. 2024. ConIQA: A deep learning method for perceptual image quality assessment with limited data. *Scientific Reports* 14, 1 (2024), 20066.

[19] Kede Ma, Wentao Liu, Kai Zhang, Zhengfang Duanmu, Zhou Wang, and Wangmeng Zuo. 2017. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing* 27, 3 (2017), 1202–1213.

[20] O. Ieremeiev N. Ponomarenko and et al. 2013. Color image database TID2013: Peculiarities and preliminary results. In *European workshop on visual information processing (EUVIP)*. IEEE, 106–111.

[21] Heeseok Oh, Jinwoo Kim, Taewan Kim, and Sanghoon Lee. 2022. Convolved quality transformer: Image quality assessment via long-range interaction between local perception. *IEEE Access* 10 (2022), 102968–102980.

[22] J. Sturtz P. Yang and Q. Letu. 2023. Progress in blind image quality assessment: a brief review. *Mathematics* 11, 12 (2023), 2766.

[23] S. Dadsetan S. Alireza Golestaneh and K. M. Kitani. 2022. No-reference image quality assessment via transformers, relative ranking, and self-consistency. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1220–1230.

[24] D. Maniry S. Bosse, T. Wiegand K.-R. Müller, and W. Samek. 2017. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on image processing* 27, 1 (2017), 206–219.

[25] Y. Gong S. Lao, T. Wu S. Shi, S. Yang, and et al. 2022. Attentions help cnns see better: Attention-based hybrid image quality assessment network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 1140–1149.

[26] Q. Yan S. Su, X. Ge Y. Zhu, C. Zhang, and et al. 2020. Blindly assess image quality in the wild guided by a self-adaptive hyper network. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3667–3676.

[27] T. Wu S. Yang, Y. Gong S. Shi, S. Lao, and et al. 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*. 1191–1200.

[28] Darwin Saire and Adin Ramirez Rivera. 2022. Global and Local Features Through Gaussian Mixture Models on Image Semantic Segmentation. *IEEE Access* 10 (2022), 77323–77336.

[29] Hamid R Sheikh and Alan C Bovik. 2005. A visual information fidelity approach to video quality assessment. In *The first international workshop on video processing and quality metrics for consumer electronics*, Vol. 7. sn, 2117–2128.

[30] H. R. Sheikh and A. C. Bovik. 2006. Image information and visual quality. *IEEE Transactions on image processing* 15, 2 (2006), 430–444.

[31] M. Nuutinen T. Virtanen, P. Oittinen M. Vaahteranoksa, and J. Häkkinen. 2014. CID2013: A database for evaluating no-reference image quality assessment algorithms. *IEEE Transactions on Image Processing* 24, 1 (2014), 390–402.

[32] K. Ma W. Zhang, D. Deng J. Yan, and Z. Wang. 2020. Blind Image Quality Assessment Using A Deep Bilinear Convolutional Neural Network. *IEEE Transactions on Circuits and Systems for Video Technology* 30, 1 (2020), 36–47.

[33] Shiqi Wang, Xiang Zhang, Siwei Ma, and Wen Gao. 2013. Reduced reference image quality assessment using entropy of primitives. In *2013 Picture Coding Symposium (PCS)*. IEEE, 193–196.

[34] Xiaoqi Wang and Yun Zhang. 2024. Global-Local Progressive Integration Network for Blind Image Quality Assessment. *arXiv preprint arXiv:2408.03885* (2024).

[35] Z. Wang and A. C. Bovik. 2002. A universal image quality index. *IEEE signal processing letters* 9, 3 (2002), 81–84.

[36] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* 13, 4 (2004), 600–612.

[37] V. Wasson and B. Kaur. 2019. Full Reference Image Quality Assessment from IQA Datasets: A Review. In *2019 6th International Conference on Computing for Sustainable Global Development (INDIACom)*. IEEE, 735–738.

[38] J. Van De Weijer X. Liu and A. D. Bagdanov. 2017. Rankiqa: Learning from rankings for no-reference image quality assessment. In *Proceedings of the IEEE international conference on computer vision*. 1040–1049.

[39] H. Zhu Y. Fang, K. Ma Y. Zeng, and Z. Wang. 2020. Perceptual quality assessment of smartphone photography. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3677–3686.

[40] Junfeng Yang, Jing Fu, Zhen Zhang, Limei Liu, Qin Li, Wei Zhang, and Wenzhi Cao. 2024. Align-IQA: aligning image quality assessment models with diverse human preferences via customizable guidance. In *Proceedings of the 32nd ACM International Conference on Multimedia*. 10008–10017.

[41] Y. K. Yun and W. Lin. 2024. You Only Train Once: A Unified Framework for Both Full-Reference and No-Reference Image Quality Assessment. arXiv:2310.09560 [cs.CV] https://arxiv.org/abs/2310.09560