scientific reports

Check for updates

OPEN

QualityNet: A multi-stream fusion framework with spatial and channel attention for blind image quality assessment

Muhammad Azeem Aslam[1,2]✉, Xu Wei[2], Hassan Khalid[3], Nisar Ahmed[4], Zhu Shuangtong[2], Xin Liu[1] & Yimei Xu[1]

This study introduces a novel Blind Image Quality Assessment (BIQA) approach leveraging a multi-stream spatial and channel attention model. Our method addresses challenges posed by diverse image content and distortions by integrating feature maps from two distinct backbones. Through spatial and channel attention mechanisms, our algorithm prioritizes regions of interest, enhancing its ability to capture crucial image details. Extensive evaluations on four benchmark datasets demonstrate superior performance compared to existing methods, closely aligning with human perceptual assessment. Our approach exhibits exceptional generalization capabilities on both authentic and synthetic distortion databases. Moreover, it demonstrates a distinctive focus on perceptual foreground information, enhancing its practical applicability. Thorough quantitative analyses underscore the algorithm's superior performance, establishing its dominance over existing methods.

Keywords Image Quality Assessment, Blind Image Quality Assessment, BIQA, Spatial Attention, Channel Attention, Multi-stream Model and objective quality assessment

In today's digital landscape, images play pivotal roles across diverse sectors such as medicine, remote sensing, and entertainment[1,2]. However, these images often suffer from quality degradation due to various factors like compression, noise, and color distortions. Such deterioration can affect visual perception or impede the extraction of crucial information[3–5], potentially affecting disease diagnosis, precision agriculture, or the overall Quality of Service (QoS) in various applications[6,7].

Addressing these challenges, Image Quality Assessment (IQA) approaches have emerged, aiming to objectively evaluate the quality of digital images[8–10]. A fundamental challenge in IQA is understanding how different distortions impact image quality and finding effective ways to mitigate their effects[9,11]. Factors such as limitations in imaging sensors, ambient conditions, transmission bandwidth, and storage space can introduce noise, blur, compression artifacts, and color distortions, all of which can obscure important image details[12]. Therefore, the development of generalized IQA algorithms becomes imperative to ensure the efficient utilization of images for their intended purposes[13].

The IQA domain holds significant implications, particularly in fields where image quality directly influences outcomes. Generalized IQA algorithms are instrumental in mitigating distortion effects, thus improving overall visual perception and enabling more accurate diagnoses and treatments in medicine, precision agriculture and enhancing QoS in entertainment contexts[6,8,14].

Objective IQA algorithms can be classified into three categories based on the access to reference information: full-reference, reduced-reference, and no-reference. Full-reference algorithms compare a distorted image to its original reference, whereas reduced-reference methods assess image quality using partial information from the original image, such as metadata or reference-extracted features[15,16]. In contrast, no-reference techniques evaluate image quality without access to the original reference, instead relying solely on the distorted image[17–19]. Figure 1 visually depicts the working principle of these three IQA approaches.

[1]School of Information Engineering, Xi'an Eurasia University, Xi'an 710065, Shaanxi, China. [2]Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China. [3]Department of Electrical Engineering, University of Engineering and Technology Lahore, Lahore 54890, Punjab, Pakistan. [4]Department of Computer Engineering, University of Engineering and Technology Lahore, Lahore 54890, Punjab, Pakistan. ✉email: azeem@eurasia.edu

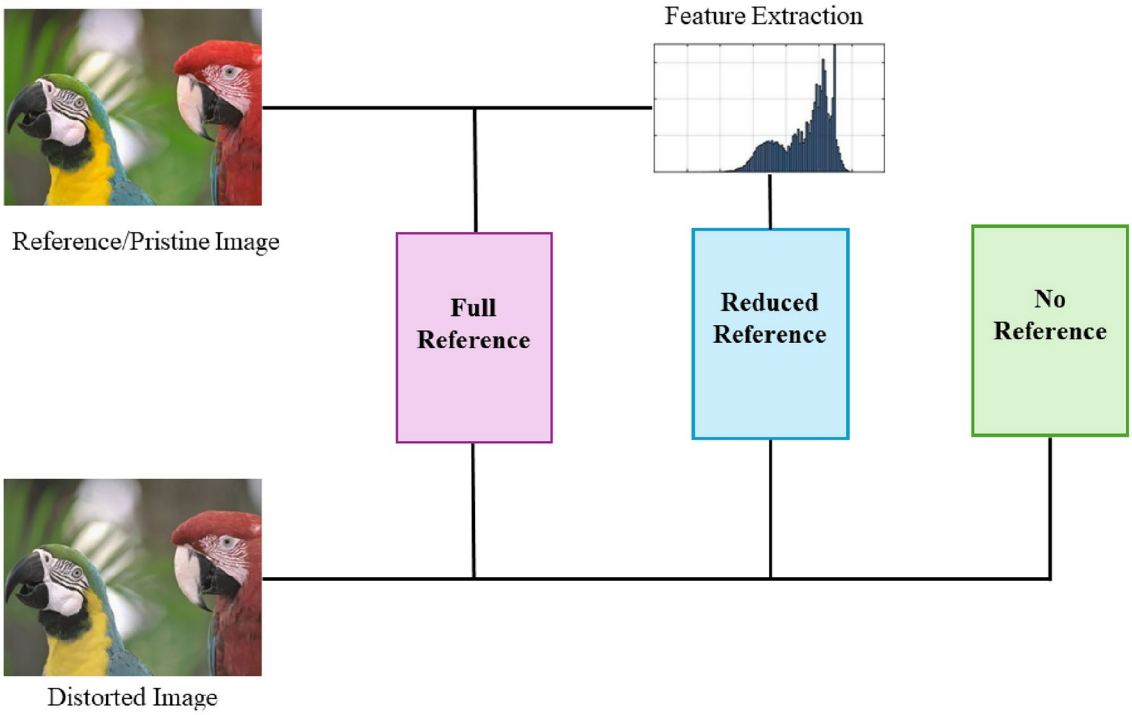


Fig. 1. Working principles of objective image quality assessment techniques

While full-reference and reduced-reference methods provide accurate assessments in controlled environments with access to reference images, they face limitations in real-world scenarios[20,21] where such references may be unavailable or impractical to obtain. In contrast, no-reference IQA methods offer a solution to this challenge by autonomously evaluating image quality without relying on reference images[22]. Therefore these methods are crucial in situations where access to reference images is limited, such as real-time video streaming[23–25], precision agriculture[3,4] or medical imaging[26,27], and various distortions found in modern digital environments[12,28].

Despite the critical importance of IQA, existing methodologies face challenges in accurately assessing image quality, particularly in scenarios where images contain multiple important regions. Traditional approaches often treat the entire image uniformly, leading to inaccuracies, especially when dealing with complex objects or natural scenes with foreground and background information.

To address these limitations, this study introduces a multi-stream spatial and channel attention algorithm. This algorithm leverages features from two backbones and integrates spatial and channel attention mechanisms to enhance predictions that closely align with human perceptual assessment. By focusing on important regions within the image, this approach improves the algorithm's capacity to capture salient image details.

Additionally, the study proposes a quality-aware loss function to enhance precision and correlation in BIQA. This loss function combines two key metrics, the Mean Absolute Error (MAE) and the Pearson Linear Correlation Coefficient (PLCC) based loss, to provide a comprehensive evaluation of image quality. By considering both the absolute error and the linear correlation between predicted and ground truth quality scores, the quality-aware loss function offers a holistic perspective on the image quality assessment process.

The effectiveness and practicality of the proposed approach are demonstrated through comprehensive evaluations of four benchmark datasets, encompassing both authentic and synthetic distortion databases. The algorithm developed in this study exhibits a notable focus on perceptual foreground information, thereby enhancing its suitability for real-world applications.

Literature review

The field of BIQA has witnessed significant advancements with the introduction of hand-engineered features-based algorithms and deep learning-based algorithms. This literature review explores and analyzes the progress made in BIQA through these two distinct approaches. The first part of the review focuses on the early stages of BIQA, where hand-engineered features played a pivotal role in developing assessment algorithms. Key characteristics, advantages, and limitations of these feature-based methods are thoroughly investigated. Subsequently, the second part of the review delves into the recent developments driven by deep learning techniques, which have revolutionized the BIQA domain. Contributions of deep learning-based algorithms are examined, highlighting their successes and addressing any challenges they might face. Through a comprehensive examination of both research streams, this review aims to provide an in-depth understanding of the evolution and current state of BIQA techniques.

Handcrafted features based approaches

Moorthy et al.[29] introduced the DIIVINE algorithm for BIQA, based on the premise that pristine images possess specific statistical properties that are altered by artifacts and distortions, leading to unnatural appearances. DIIVINE operates in two stages: the first stage identifies distortions, while the second assigns a distortion-specific quality score. Similarly, Saad et al.[30] proposed BLIINDS-II, a no-reference BIQA algorithm that assumes original images have distinct statistical properties that are modified when distortions occur. This algorithm also functions in two stages: detecting the presence of distortions and evaluating image quality based on the identified distortions. When assessed on the LIVE IQA dataset[16], BLIINDS-II demonstrated a strong correlation with human subjective scores.

Expanding on these methods, Xue et al.[31] introduced the quality-aware clustering approach for BIQA, showing comparable performance in terms of correlation with subjective scores. In further work, Xue et al.[32] developed a model using two local contrast features: the gradient magnitude map and the Laplacian of Gaussian response. This model was evaluated across three major benchmark databases, where it achieved state-of-the-art performance compared to other models, including those employing full-reference assessments. Continuing this trend, Zhang et al.[33] proposed an opinion-unaware BIQA method that does not rely on human subjective scores as ground truth. Utilizing a multivariate Gaussian model with image patches, the approach leveraged the Gaussian distribution of locally normalized luminances and the Weibull distribution of gradient magnitudes. Although this model showed reasonable performance, it was outperformed by models trained on subjectively scored images.

Addressing the need for BIQA methods targeting poor contrast images, Fang et al.[34] developed an approach using Natural Scene Statistics (NSS) features and support vector regression to predict image quality based on subjective evaluation scores. The model has been evaluated on three benchmark datasets, however, the performance was highest on the CID2013 dataset[35]. Building on the use of high-order statistics, Xu et al.[36] proposed an approach called High-Order Statistics Aggregation (HOSA), achieving competitive performance against state-of-the-art methods while requiring only a small codebook. This model was evaluated on ten different image databases with both simulated and realistic distortions, effectively addressing the challenges faced by previous feature learning-based BIQA methods.

Focusing on practical applications, Ghadiyaram et al.[37] presented a "bag of feature maps" approach, using NSS to extract non-distortion-specific features for blind image quality assessment. They tested their model on legacy databases with authentically distorted images and a new distortion-realistic database named "LIVE In the Wild Image Quality Challenge Database"[38], where it showed superior performance over existing algorithms. Similarly, Kundu et al.[39] introduced a non-reference IQA model for high dynamic range (HDR) images, combining standard and novel HDR-space features based on bandpass and differential NSS information, which resulted in optimal performance on both HDR and standard image datasets.

Sadiq et al.[40] advanced the field by proposing a BIQA technique that extracts features from both spatial and transform domains, integrating morphological gradient, discrete Laplacian, and stationary wavelet transform. The features were normalized using an adaptive joint normalization framework, resulting in superior performance compared to state-of-the-art BIQA and full-reference IQA techniques across five legacy databases. Ahmed et al.[41] followed this with the introduction of the Perceptual Image Quality Index (PIQI), combining contrast-normalized products, luminance, and gradient statistics to evaluate image quality. PIQI outperformed twelve state-of-the-art methods across six benchmark datasets in terms of RMSE, Pearson, and Spearman's correlation coefficients.

Khalid et al.[11] introduced GPR-BIQA, a novel Gaussian process-based algorithm for blind image quality assessment (BIQA). The authors developed an integrated feature selection framework tailored specifically for the BIQA problem, employing an innovative method to consolidate features by incorporating optimal characteristics from transform, spatial, and various other domains. Additionally, the algorithm utilized a Gaussian process-based regression model designed to predict image quality without focusing on specific types of distortions. The efficacy of GPR-BIQA was validated through comprehensive evaluations on both natural and synthetically distorted image databases, where it outperformed contemporary NSS and deep learning-based methodologies.

Additionally, to address image quality assessment for tone mapping algorithms in high dynamic range (HDR) images, Alotaibi et al.[42] proposed a new algorithm. This algorithm utilized sixteen distinct features and was evaluated against twenty-four existing IQA metrics, demonstrating its effectiveness on both an existing and a newly proposed dataset for the problem. The proposed approach outperformed existing tone mapping algorithms and ranked second highest on standard legacy datasets.

Deep learning-based approaches

In addition to algorithms based on NSS handcrafted features, deep learning-based methods have also emerged as effective solutions in BIQA. Gu et al.[43] introduced DIQI, a deep neural network-based algorithm capable of capturing complex image attributes, outperforming classical full-reference and state-of-the-art reduced and no-reference IQA algorithms on the TID2013 database. Similarly, Fu et al.[44] proposed a CNN-based BIQA method that achieved optimal performance on the LIVE IQA database. Bianco et al.[45] further developed DeepBIQ, which outperformed other state-of-the-art methods across several legacy benchmark datasets, including LIVE[16], CSIQ[46], TID2008[47], and TID2013[48].

Ma et al.[21] proposed MEON, a multi-task end-to-end optimized deep neural network for BIQA that consists of two sub-networks for distortion identification and quality prediction. This model demonstrated competitiveness against state-of-the-art BIQA models when tested on multiple IQA datasets. In another development, Zhang et al.[9] proposed a deep bilinear model for blind image quality assessment, addressing both synthetic and authentic distortions using two streams of deep CNNs. Ahmed et al. presented a CNN-based BIQA algorithm that utilizes an ensemble technique, showing performance advantages in BIQA. Later, Ahmed et al.[49] proposed a hybrid

BIQA method combining handcrafted and deep extracted features, which outperformed existing algorithms on seven benchmark datasets in terms of correlation with human opinion measures.

Ying et al.[50] proposed PaQ-2-PiQ, an extensive subjective picture quality database, leading to advanced models capable of predicting both global and local image quality accurately. This significant development helps address the challenges of blind perceptual image quality assessment in real-world scenarios. Similarly, Hosu et al.[17] introduced the KonIQ-10k dataset, one of the largest IQA datasets available, along with KonCept512, a deep learning-based model that exhibited excellent generalization characteristics compared to state-of-the-art algorithms. Building on deep learning techniques, Ahmed et al.[51] proposed an IQA model using activations of pre-trained deep neural architectures as features from an ensemble of Gaussian process regression models, achieving state-of-the-art performance across various datasets. On the same lines, Varga[52] proposed an IQA algorithm with multiple neural architecture-based decision feature fusion. The algorithm has outperformed several IQA algorithms, however, the computationally and memory efficacy is not optimistic.

Zhang et al.[1] proposed a method for continual learning in BIQA, in which the model adapts incrementally from a stream of IQA datasets, demonstrating superior performance over traditional training techniques. Wang et al.[6] investigated the application of Contrastive Language-Image Pre-training (CLIP) models, emphasizing that the extensive visual-linguistic priors embedded within CLIP can effectively evaluate image quality perception without the necessity of task-specific training. Their method achieved an approximate accuracy of 80% across all five attributes evaluated. Subsequently, Sang et al.[53] introduced a self-supervised learning algorithm aimed at addressing the poor generalizability observed in deep learning-based IQA methods. A significant advantage of this algorithm is its compatibility with deployment on edge devices, achieving a performance level surpassing that of the teacher network.

Li et al.[54] proposed a CNN-based BIQA algorithm incorporating content-awareness and distortion-sensitivity mechanisms, which exhibited superior performance across various image and distortion types. Yang et al.[55] introduced a transformer-based BIQA algorithm that demonstrated higher performance compared to other IQA methods. However, it was noted that their algorithm requires more computational resources, yielding only marginal improvements over conventional state-of-the-art IQA methods.

The existing literature presents a comprehensive spectrum of deep learning-based BIQA algorithms, illustrating their efficacy and potential in delivering precise image quality assessments. Despite these advancements, techniques such as transformer-based IQA necessitate large-scale datasets to prevent overfitting and demand substantial GPU memory along with expensive hardware for inference during later stages. Therefore, the development of an algorithm featuring robust feature fusion, effective attention mechanisms, and enhanced generalizability across IQA datasets is imperative for future research.

Proposed model

A detailed visual representation of the proposed architecture is reported in Figure 2a. The model consists of two parallel backbones inspired by Resnet and EfficentNet that are optimized for the task using transfer learning to ensure robust deep feature volume extraction. The feature volume extracted from each backbone is then reduced in size through Global Average Pooling to capture global information in a regularized manner. Spatial and channel attention modules are then applied in each parallel pipeline followed by feature fusion, and finally, an optimized feed-forward neural regression head is added for quality assessment. The spatial and channel attention mechanisms are inspired by MIRNET[56] and the final architectures used for both of these attention modules are depicted in Figure 2b and 2c.

The spatial attention block highlights the most important parts of an image by learning individual weights for each pixel. This allows the model to focus on specific regions and suppress irrelevant ones, ultimately enhancing the quality assessment process. On the other hand, the channel attention block emphasizes the significant channels within a convolutional layer by learning weights that highlight essential channels while suppressing less relevant ones. This approach enables the model to capture and leverage the most vital information from each channel, leading to a more effective overall feature representation. By integrating the spatial and channel attention blocks into the final model, the model's ability to discern critical image details is enhanced, thus improving image quality assessment performance.

Spatial block

The spatial attention block incorporates several essential operations to highlight important regions within an image. It utilizes both average and maximum pooling operations to extract significant features. Average pooling calculates the mean value across a specific area, while maximum pooling selects the highest value. These operations help in downsampling and abstracting feature maps, making it easier to identify important spatial information.

Mathematically, the average pooling for spatial attention is expressed as:

$$avg\_pool[i,j,k] = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} x[i,h,w,k] \tag{1}$$

Moreover, maximum pooling for spatial attention is represented as:

$$max\_pool[i,j,k] = \max_{h=1}^{H} \left(\max_{w=1}^{W} x[i,h,w,k] \right) \tag{2}$$

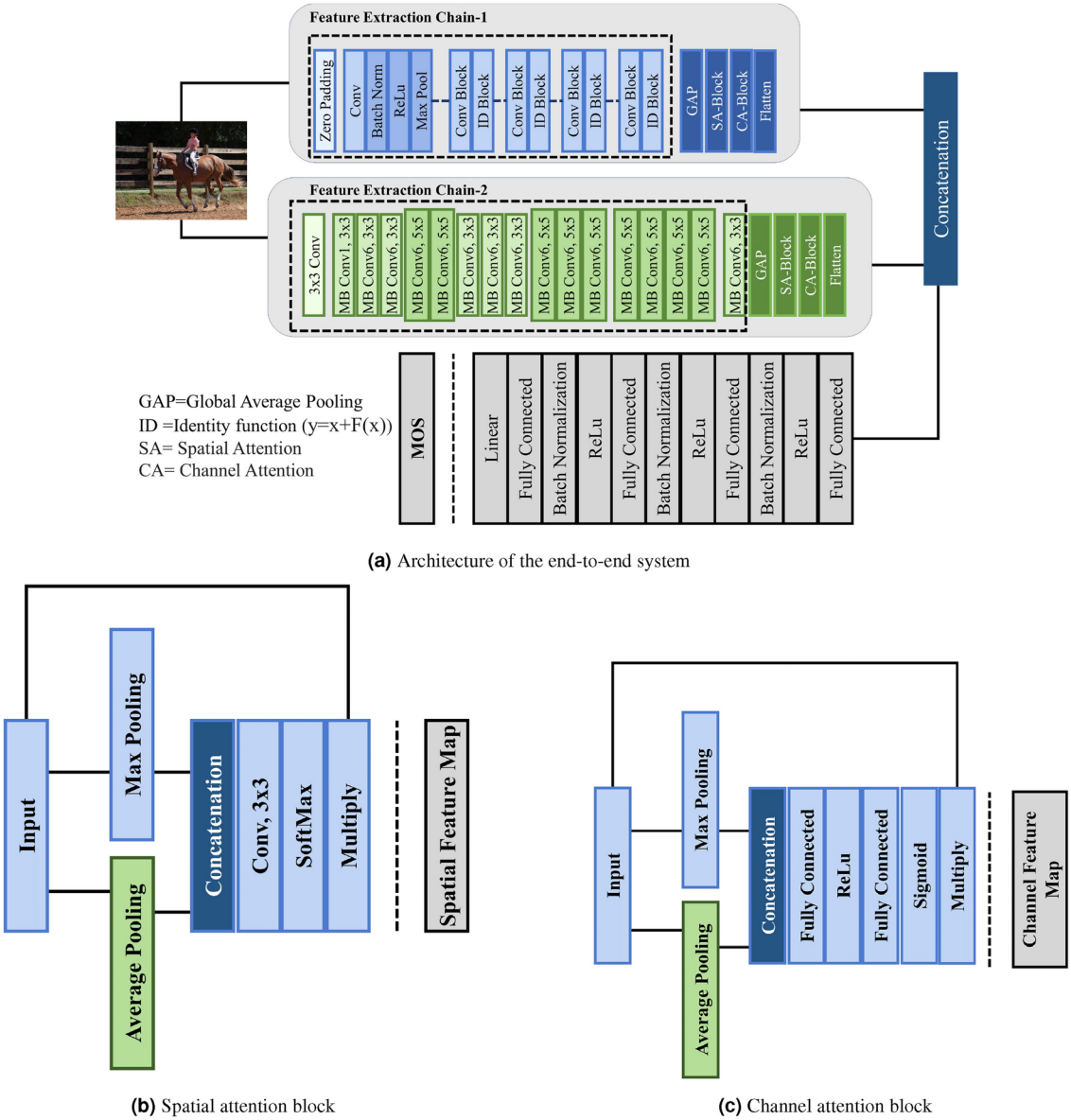


Fig. 2. Illustrations of different components of the system

The outputs from the pooling operations are concatenated along the last dimension, preserving spatial information while combining features extracted through different methods. This concatenation enhances the model's ability to capture diverse aspects of the input data.

Mathematically, the operation can be denoted as:

$$concat[i, j, k] = [avg\_pool[i, j, k], max\_pool[i, j, k]] \tag{3}$$

Following concatenation, a convolutional layer is applied to the concatenated tensor. This layer utilizes a 3×3 kernel to convolve over the spatial dimensions, producing an output tensor with a singular channel. The sigmoid activation function is then applied to generate a probabilistic spatial representation.

Mathematically, the operation is expressed as:

$$\text{conv}[i, j, k] = \sigma\left(\sum_{m=-1}^{1}\sum_{n=-1}^{1} w[m, n] \cdot \text{concat}[i, j + m, k + n] + b\right) \tag{4}$$

where:

- w represents the convolution kernel,
- σ denotes the sigmoid activation function

Finally, the spatial feature map is generated through element-wise multiplication between the input tensor and the output tensor obtained from the convolutional layer. This operation emphasizes regions of interest identified by the spatial attention mechanism.

Mathematically, the operation is given by:

$$\text{Spatial Feature Map}[i, j, k] = x[i, j, k] \cdot \text{conv}[i, j, k] \tag{5}$$

These combined operations within the spatial block enable the model to focus on relevant spatial features within the input image effectively.

Channel attention block

The channel attention mechanism serves the purpose of emphasizing important channels within the input feature maps while suppressing less relevant ones. It takes an input tensor x and performs the following operations:

First, it applies average pooling along the spatial dimensions of the input tensor (height and width). This operation facilitates the adaptive weighting of channel importance, contributing to refined feature representation and learning.

Mathematically, it can be expressed as:

$$\text{avg\_pool}[i, 1, 1, k] = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} x[i, h, w, k] \tag{6}$$

Second, it employs max-pooling along the spatial dimensions of the input tensor. This operation captures the maximum activation intensity of each channel across the comprehensive spatial expanse of the feature map. The utilization of max-pooling in this context facilitates the adaptive weighting of channel importance, contributing to refined feature representation and learning.

Mathematically, it can be expressed as:

$$\text{max\_pool}[i, 1, 1, k] = \max_{h=1}^{H} \left(\max_{w=1}^{W} x[i, h, w, k] \right) \tag{7}$$

Third, it concatenates the output tensors from the average pooling and maximum pooling operations. This results in an output tensor with the shape $(\text{batch\_size}, 1, 1, 2 \times \text{channels})$.

Mathematically, it can be expressed as:

$$\text{concat}[i, 1, 1, k] = [\text{avg\_pool}[i, 1, 1, k], \text{max\_pool}[i, 1, 1, k]] \tag{8}$$

Fourth, it applies a fully connected layer called fc1 to reduce the dimensionality of the concatenated tensor. It introduces non-linearity through the rectified linear unit (ReLU) activation function. After fc1, another fully connected layer called fc2 is introduced to restore the dimensionality to the original size of the last dimension of the input tensor x. The sigmoid activation function is applied for the adaptive modulation of each unit within the output. This modulation is pivotal, as it enables the model to dynamically weight the importance of each channel within the feature map.

Mathematically, it can be expressed as:

$$\text{Fcq}[i, 1, 1, k] = \text{ReLU} \left(\sum_{j=1}^{2 \times \text{channels}} W_{\text{Fc1}}[j, k] \cdot \text{concat}[i, 1, 1, j] + b_{\text{Fc1}}[k] \right) \tag{9}$$

$$\text{Fc2}[i, 1, 1, k] = \text{Sigmoid} \left(\sum_{j=1}^{\text{channels}} W_{\text{Fc2}}[j, k] \cdot \text{fc1}[i, 1, 1, j] + b_{\text{Fc2}}[k] \right) \tag{10}$$

Finally, it performs an element-wise multiplication operation between the input tensor x and the output tensor obtained after the fc2 layer. This operation acts as a gate, integrating the learned channel-wise information from dense2 back into the original input tensor. Consequently, the network gains the capacity to selectively amplify or attenuate specific channels based on their learned relevance, enhancing the model's discriminatory power and feature representation capabilities. Mathematically, it can be expressed as:

$$\textit{Channel Attention Feature Map}[i, j, k] = x[i, j, k] \cdot \text{Fc2}[i, j, k] \tag{11}$$

Final proposed model

The complete architecture of the proposed model can be represented mathematically as follows:

Input layer

The input tensor is denoted as x.

Chain-1

The chain-1 takes the input tensor x and performs a series of operations:

$$x_2 = \text{Feature Extraction Chain1}(x) \tag{12}$$

$$x_2[i,j,k] = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} x_2[i,h,w,k] \tag{13}$$

$$x_2[i,j,k] = \text{Reshape}\left((1,1,-1)\right)\left(x_2[i,j,k]\right) \tag{14}$$

$$x_2 = \text{spatial\_attention}(x_2) \tag{15}$$

$$x_2 = \text{channel\_attention}(x_2) \tag{16}$$

$$x_2[i,j,k] = \text{Flatten}\left(\sum_{h=1}^{H} \sum_{w=1}^{W} x_2[i,h,w,k]\right) \tag{17}$$

Chain-2

The chain-2 takes the input tensor x and performs a series of operations, here the Feature Extraction Chain 2 is derived from EfficentNetB7 :

$$x_3 = \text{Feature Extraction Chain2}(x) \tag{18}$$

$$x_3[i,j,k] = \frac{1}{H \times W} \sum_{h=1}^{H} \sum_{w=1}^{W} x_3[i,h,w,k] \tag{19}$$

$$x_3[i,j,k] = \text{Reshape}\left((1,1,-1)\right)\left(x_3[i,j,k]\right) \tag{20}$$

$$x_3 = \text{spatial\_attention}(x_3) \tag{21}$$

$$x_3 = \text{channel\_attention}(x_3) \tag{22}$$

$$x_3[i,j,k] = \text{Flatten}\left(\sum_{h=1}^{H} \sum_{w=1}^{W} x_3[i,h,w,k]\right) \tag{23}$$

Feature fusion

The outputs from the Feature Extraction Chain-1 and Feature Extraction Chain-2 are concatenated to form a single tensor:

$$\text{concatenated\_output}[i,j,k] = \text{Concatenate}()([x_2[i,j,k], x_3[i,j,k]]) \tag{24}$$

Regression head

The concatenated output tensor mentioned in 24 then undergoes several fully connected layers with batch normalization and dropout applied:

$$x[i,j,k] = \max\left(0, \sum_{m=1}^{1024} W_1[m,k] \cdot \text{concatenated\_output}[i,j,m] + b_1[k]\right) \tag{25}$$

- W_1 is the weight matrix for the first dense layer.
- b_1 is the bias term for the first dense layer.

$$x[i,j,k] = \frac{x[i,j,k] - \mu_k}{\sigma_k + \epsilon} \tag{26}$$

- μ_k is the mean for the batch normalization of the first dense layer.
- σ_k is the standard deviation for the batch normalization of the first dense layer.

- ϵ is a small constant for numerical stability.

$$x[i, j, k] = x[i, j, k] \cdot \text{Bernoulli}(p = 0.75) \tag{27}$$

- $\text{Bernoulli}(p = 0.75)$ is a random variable with a Bernoulli distribution and probability $p = 0.75$.

$$x[i, j, k] = \max\left(0, \sum_{m=1}^{512} W_2[m, k] \cdot x[i, j, m] + b_2[k]\right) \tag{28}$$

- W_2 is the weight matrix for the second dense layer.
- b_2 is the bias term for the second dense layer.

$$x[i, j, k] = \frac{x[i, j, k] - \mu_k}{\sigma_k + \epsilon} \tag{29}$$

- μ_k is the mean for the batch normalization of the second dense layer.
- σ_k is the standard deviation for the batch normalization of the second dense layer.
- ϵ is a small constant for numerical stability.

$$x[i, j, k] = x[i, j, k] \cdot \text{Bernoulli}(p = 0.75) \tag{30}$$

- $\text{Bernoulli}(p = 0.75)$ is a random variable with a Bernoulli distribution and probability $p = 0.75$.

$$x[i, j, k] = \max\left(0, \sum_{m=1}^{256} W_3[m, k] \cdot x[i, j, m] + b_3[k]\right) \tag{31}$$

- W_3 is the weight matrix for the third dense layer.
- b_3 is the bias term for the third dense layer.

$$x[i, j, k] = \frac{x[i, j, k] - \mu_k}{\sigma_k + \epsilon} \tag{32}$$

- μ_k is the mean for the batch normalization of the third dense layer.
- σ_k is the standard deviation for the batch normalization of the third dense layer.
- ϵ is a small constant for numerical stability.

$$x[i, j, k] = x[i, j, k] \cdot \text{Bernoulli}(p = 0.5) \tag{33}$$

- $\text{Bernoulli}(p = 0.5)$ is a random variable with a Bernoulli distribution and probability $p = 0.5$.

The final stage, predicts the blind image quality using a dense layer with a linear activation function. This layer takes the refined feature maps as input and directly outputs a numerical quality score, providing an objective assessment of the image's perceptual quality.

$$\text{predictions}[i] = \sum_{j=1}^{\text{num\_classes}} W_{\text{predictions}}[j] \cdot x[i, j] + b_{\text{predictions}}[j] \tag{34}$$

- predictions is the vector of final predictions.
- $W_{\text{predictions}}$ is the weight matrix for the output layer.
- $b_{\text{predictions}}$ is the bias term for the output layer.
- $x[i, j]$ represents the j-th element of the vector x at batch i.
- The linear activation function is applied to obtain the final predictions.

Dataset	Type	Total Images	Resolution	Score Range
Koniq-10k	Authentically distorted	10,073	Multiple	0–100
BIQ2021	Authentically distorted	12,000	512×512	0–1
LIVEC	Authentically distorted	1,162	500×500	0–100
TID2013	Synthetically distorted	3,000	512×384	0–9

Table 1. Summary of Image Quality Assessment Datasets.

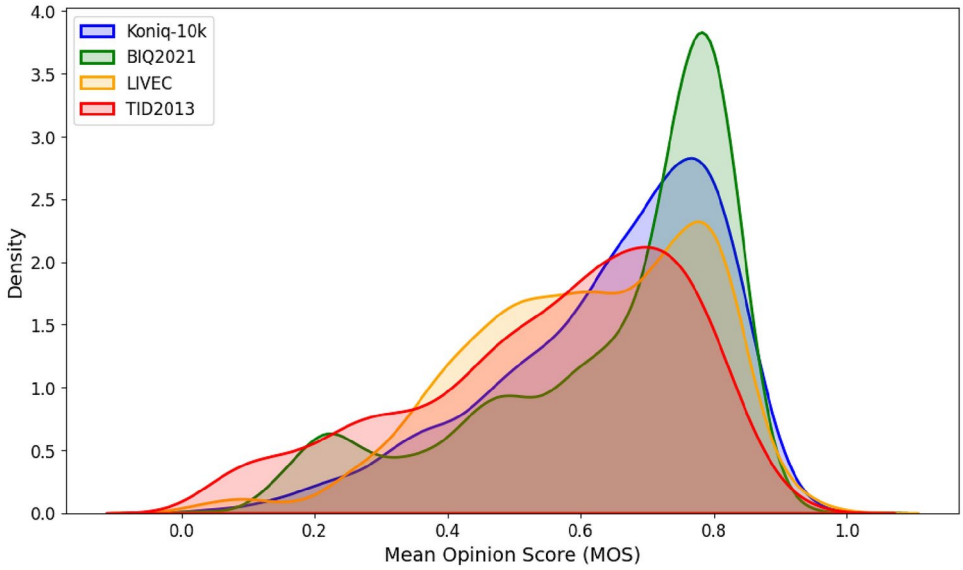


Fig. 3. Working principles of objective image quality assessment techniques

Experimental results

Datasets

The selection of relevant datasets is crucial to thoroughly assessing BIQA models. A comprehensive assessment is conducted using both synthetic and authentic distortion image datasets. The synthetic distortion image dataset allows for the evaluation of model performance under controlled or laboratory conditions, resulting in valuable preliminary insights. Meanwhile, authentic distortion image datasets make it easier to evaluate model performance in real-world scenarios, ensuring that the results are relevant and reliable. The use of these diverse datasets allows for a comprehensive evaluation of BIQA models across various contexts.

Among the datasets used in this study is the TID2013 dataset[48], which contains 25 reference images that are distorted by simulating 24 types of distortions at five granularities. The LIVE in the Wild Challenge dataset[38], which consists of 1,161 images with varying degrees and types of distortion. The images included in this dataset are captured by different and diverse image-capturing devices, so that the natural artifacts and distortions may be authentically captured. The KonIQ-10K[17] dataset contains 10,073 authentically distorted images, the dataset is one of the most diverse and large-scale datasets available for the design and evaluation of the image quality assessment algorithm in the natural environment. The BIQ2021[10] is the largest dataset of authentically distorted images and contains 12,000 where the ground truth quality score for the dataset is provided in the form of MOS. The summary for these datasets is tabulated in Table 1. Quality score normalized distribution plots for different datasets is reported in Figure3.

Evaluation metrics

To validate and compare the performance of the designed algorithm, standard evaluation metrics commonly employed for typical regression problems are utilized. These metrics provide essential insights into the algorithm's accuracy and effectiveness in predicting continuous quality scores. Further details on the specific evaluation metrics and their application in the context of blind image quality assessment will be discussed in the subsequent sections.

Pearson Linear Correlation Coefficient (PLCC)

The Pearson correlation coefficient is a widely used metric that quantifies the strength of a linear relationship between variables. In the context of image quality assessment, it serves as a common evaluation measure to assess the alignment between a BIQA model's predicted quality scores and the ground truth quality scores. A high Pearson correlation coefficient indicates a strong linear relationship, suggesting that the model's predictions

closely match the actual quality scores, thereby validating its effectiveness in objective image quality assessment. The PLCC can be given by 35:

$$\mathrm{PLCC} = \frac{\mathrm{Cov}(Y, Y')}{\sqrt{\mathrm{Var}(Y)\mathrm{Var}(Y')}} \tag{35}$$

The value of PLCC ranges from -1 to +1. A PLCC value of +1 indicates a perfect positive linear correlation, meaning that as the ground truth quality scores increase, the BIQA model's predicted scores also increase proportionally. A PLCC value of -1 indicates a perfect negative linear correlation, meaning that as the ground truth quality scores increase, the BIQA model's predicted scores decrease proportionally. Finally, a PLCC value of 0 indicates that there is no linear correlation between the two sets of scores.

Spearman ranked order correlation coefficient (SROCC)

The SROCC is a valuable metric for measuring the strength of a monotonic relationship between two variables. In the context of image quality assessment, it is employed to compare a BIQA model's predicted quality scores with the ground truth quality scores, without considering the exact magnitudes of the scores. SROCC assesses the consistency in the ranking order of scores, providing insights into how well the model captures the relative image quality judgments, irrespective of specific numerical values. This makes SROCC particularly useful when evaluating the model's ability to rank images according to their perceptual quality accurately. The SROCC can be given by 36:

$$\mathrm{SROCC} = 1 - \frac{6\sum D^2}{n(n^2-1)} \tag{36}$$

where D is the difference between the ranks of Y and Y', and n is the number of samples. The value of SROCC ranges between -1 and +1. An SROCC value of +1 indicates a perfect monotonic correlation, meaning that as the ground truth quality scores increase (or decrease), the BIQA model's predicted scores also increase (or decrease). An SROCC value of -1 indicates a perfect negative monotonic correlation, meaning that as the ground truth quality scores increase (or decrease), the BIQA model's predicted scores decrease (or increase). Finally, an SROCC value of 0 indicates no monotonic correlation between the two sets of scores.

Kendall ranked order correlation coefficient (KROCC)

Kendall's Tau or KROCC is another valuable metric utilized to quantify the correlation between the predicted and ground truth variables, particularly in the context of image quality assessment. This measure focuses on evaluating the similarity of rankings between the two sets of variables. By assessing the concordant and discordant pairs of rankings, Kendall's Tau provides valuable insights into how well the BIQA model captures the relative image quality judgments, regardless of the actual numerical values of the quality scores. Its use is especially pertinent when evaluating the model's ability to maintain consistent ranking orders between the predicted and ground truth quality scores, making it a robust tool for assessing the model's performance in terms of perceptual quality assessment. The KROCC can be given by 37:

$$\tau = \frac{2P}{n(n-1)} \tag{37}$$

where P is the number of concordant pairs minus the number of discordant pairs, and n is the number of observations. A concordant pair is a pair of observations that have the same order in both variables being compared (i.e., they are either both greater than or both less than each other). A discordant pair is a pair of observations that have opposite orders in the two variables (i.e., one is greater than the other in one variable, but smaller in the other variable). The Kendall tau coefficient (denoted by the symbol τ) ranges from -1 to +1, with values of -1 indicating a perfect negative association, 0 indicating no association, and +1 indicating a perfect positive association.

Implementation details

The proposed algorithm is trained and evaluated on an NVIDIA P5000 graphics card with a memory bandwidth of 16 GB. The optimizer used for training is Adam, with an initial learning rate of $1e-4$. If there is no improvement in validation performance for two consecutive epochs, the learning rate decreases. The minimum learning rate is set to $1e-8$. To prevent overfitting, an early stopping criterion is implemented. The total number of training epochs is set to 100, and the batch size is 10. The model's image dimensions are 224×224.

For training and testing, each dataset is split into three non-overlapping parts: training, validation, and testing. These splits are used for training and evaluating the proposed model and comparing it with existing best-performing models. All algorithms are evaluated on the same image size and split to ensure a fair comparison, providing a standardized evaluation environment for accurate and unbiased comparisons. The training curve for the algorithm is reported in Figure4.

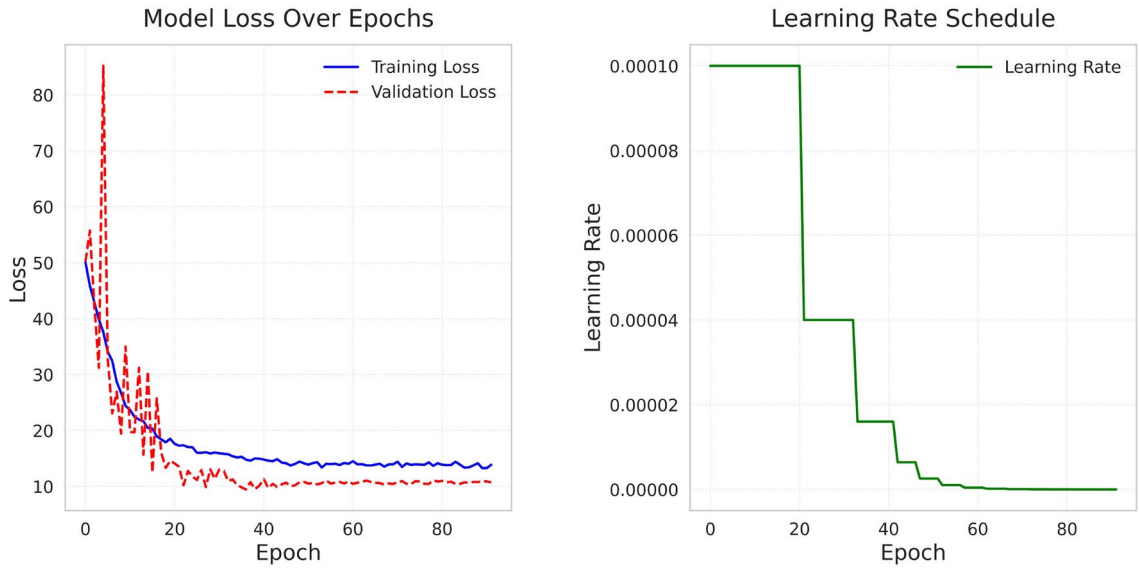


Fig. 4. Traning and Learning rate scheduler curves

Loss function

In the training of the proposed model, a modified loss function is introduced to capture better correlation and absolute value performance. This modified loss function is designed to include both MAE and PLCC terms.

The MAE term in the loss function emphasizes the absolute difference between the predicted and ground truth quality scores. By minimizing this term, the model aims to reduce the average magnitude of errors, ensuring accurate predictions regarding absolute value. On the other hand, the PLCC term in the loss function focuses on capturing the linear relationship between the predicted and ground truth quality scores. A higher PLCC value indicates a stronger linear correlation, demonstrating the model's capability to align its predictions with human perceptual judgments.

The proposed model optimizes for accurate absolute value predictions and strong linear correlation with human judgment by combining MAE and PLCC terms in the loss function. This modification enables the model to effectively balance between precise quality score predictions and maintaining a close alignment with human perception, leading to enhanced performance in blind image quality assessment. The loss function can be implemented by using the equation 38.

$$\text{Loss} = \lambda_1 \text{MAE} + \lambda_2 \text{PLCC}_{\text{Loss}} \tag{38}$$

where:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |Y_i - \hat{Y}_i| \tag{39}$$

$$\text{PLCC}_{\text{Loss}} = 1 - \frac{\text{Cov}(Y, \hat{Y})}{\sqrt{\text{Var}(Y)\text{Var}(\hat{Y})}} \tag{40}$$

where Y_i is the ground truth score, $\hat{Y}_i$ is the predicted score, and n is the number of samples. Here, $\lambda_1 = 1$ and $\lambda_2 = 10$. The weight terms assigned to each component of the loss function is estimated using parametric analysis. The significance of using a combined MAE and PLCC loss for regression tasks lies in its robustness to outliers, assessment of linear relationships, balanced evaluation of accuracy and correlation, customization of performance based on task requirements, and potential enhancement of model generalizability. The values for the λ_1 and λ_2 are determined using the grid search, the plot for the search is reported in Figure 5.

Performance evaluation

BIQA deals with the assessment of image quality in the absence of reference information and therefore the model evaluation is performed via quantitative and qualitative analysis which are discussed further.

Quantitative analysis

The quantitative analysis is performed by focusing on training and evaluating the proposed model on authentically distorted datasets and one synthetically distorted dataset. The quantitative performance of the proposed model

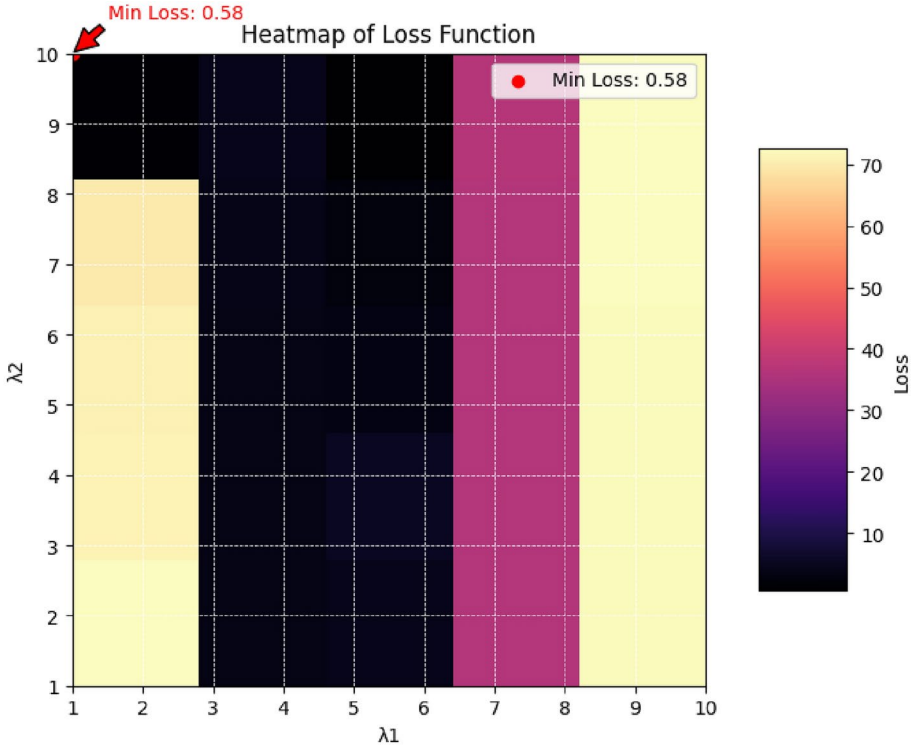


Fig. 5. The grid search for λ_1 and λ_2 values selection

Approach	PLCC	SROCC	KROCC
NIQE[58]	0.37	0.31	0.21
CNNIQA[18]	0.39	0.16	0.11
BRISQUE[59]	0.43	0.37	0.26
PI[60]	0.45	0.34	0.24
NRQM[61]	0.46	0.33	0.23
ILNIQE[33]	0.52	0.49	0.34
DBCNN[9]	0.55	0.43	0.31
PAQ2PIQ[50]	0.58	0.40	0.28
CLIPIQA_VITL14_512[6]	0.61	0.53	0.37
CLIPIQA[6]	0.65	0.58	0.41
MUSIQ-KONIQ[19]	0.68	0.58	0.41
CLIPIQA_RN50_512[6]	0.69	0.59	0.42
MANIQA[62]	0.69	0.59	0.42
CLIPIQA+[6]	0.70	0.63	0.45
GPR-BIQA[11]	*0.91*	*0.90*	*0.73*
Proposed Method	**0.92**	**0.93**	**0.77**

Table 2. Performance evaluation of various methods on TID2013 dataset: Boldface and italic text refer to the best and runner-up methods.

is compared with well-known existing approaches. Although additional datasets are available, the evaluation primarily focuses on the largest datasets in the domain.

Correlation analysis

To perform a thorough evaluation of the proposed image quality assessment model correlation analysis is conducted using PLCC, SROCC, and KROCC. These correlation measures are useful for evaluating image quality assessment performance in comparison to subjective evaluation by humans.

In this context, Table 2 reports the findings for the TID2013 dataset. It can be noted that among all the compared approaches, the proposed approach outperformed. However, given the complexity of distortions and artifacts in real-world images, models trained solely on synthetic distortion datasets may not accurately reflect

Algorithm Name	PLCC	SROCC	KROCC
BRISQUE[59]	0.21	0.23	0.15
NRQM[61]	0.48	0.37	0.25
NIQE[58]	0.32	0.38	0.26
PI[60]	0.47	0.46	0.31
ILNIQE[33]	0.52	0.55	0.39
PAQ2PIQ[50]	0.71	0.64	0.46
MANIQA[62]	0.72	0.66	0.47
CLIPIQA[6]	0.72	0.66	0.47
GPR-BIQA[11]	0.71	0.68	0.52
CNNIQA[18]	0.80	0.76	0.56
CLIPIQA+[6]	0.85	0.80	0.61
DBCNN[9]	0.86	0.84	0.66
CLIPIQA_VITL14_512[6]	0.87	0.84	0.67
MUSIQ-KONIQ[19]	0.85	0.84	0.65
CLIPIQA_RN50_512[6]	*0.86*	*0.84*	*0.67*
Proposed Method	**0.89**	**0.88**	**0.69**

Table 3. Performance evaluation of various methods on KonIQ-10K dataset: Boldface and italic text refer to the best and runner-up method.

Algorithm Name	PLCC	SROCC	KROCC
NRQM[61]	0.41	0.30	0.20
BRISQUE[59]	0.35	0.31	0.21
ILNIQE[33]	0.49	0.44	0.30
NIQE[58]	0.48	0.45	0.31
PI[60]	0.52	0.46	0.31
CNNIQA[18]	0.63	0.61	0.43
MANIQA[62]	0.72	0.66	0.47
GPR-BIQA[11]	0.66	0.66	0.46
CLIPIQA[6]	0.69	0.70	0.51
PAQ2PIQ[50]	0.75	0.72	0.53
DBCNN[9]	0.79	0.76	0.57
CLIPIQA_VITL14_512[6]	0.77	0.77	0.57
MUSIQ-KONIQ[19]	0.83	0.79	0.60
CLIPIQA+[6]	0.83	0.80	0.61
CLIPIQA_RN50_512[6]	*0.82*	*0.82*	*0.62*
Proposed Method	**0.86**	**0.84**	**0.68**

Table 4. Performance evaluation of various methods on LiveC dataset: Boldface and italic text refer to the best and runner-up methods.

quality estimation. Therefore, thorough testing and evaluation are conducted on the three largest authentically distorted image datasets to ensure robust performance.

The tables 3, 4, and 5 present the results for the KonIQ-10k[17], LiveC[57], and BIQ2021[10] datasets respectively. A visual representation of the performance of various algorithms is shown in Figure 6.

In the field of BIQA, the accuracy of the correlation between predicted and authentic ground-truth scores is crucial to emulate human perceptual judgment. The proposed approach outperforms the leading evaluation metrics, as evidenced by higher PLCC, SROCC, and KROCC scores.

Explained variance score (EVS)

The EVS is a metric that quantifies the proportion of variance in the dependent variable that the model explains. It is similar to R-squared but can have values below 0 in cases where the model performs worse than a simple mean prediction. It is calculated as:

1. Compute the total sum of squares (TSS): This represents the total variance in the dependent variable.

Algorithm Name	PLCC	SROCC	KROCC
ILNIQE[33]	0.28	0.26	0.20
NIQE[58]	0.30	0.27	0.31
NRQM[61]	0.42	0.30	0.21
PI[60]	0.53	0.46	0.32
BRISQUE[59]	0.70	0.60	0.31
CNNIQA[18]	0.64	0.61	0.43
MANIQA[62]	0.73	0.66	0.48
GPR-BIQA[11]	0.66	0.66	0.46
CLIPIQA[6]	0.69	0.70	0.51
PAQ2PIQ[50]	0.76	0.72	0.54
DBCNN[9]	0.79	0.76	0.57
CLIPIQA_VITL14_512[6]	0.77	0.77	0.58
MUSIQ-KONIQ[19]	0.83	0.79	0.60
CLIPIQA+[6]	0.84	0.81	0.62
CLIPIQA_RN50_512[6]	*0.82*	*0.82*	*0.62*
Proposed Method	**0.85**	**0.86**	**0.67**

Table 5. Performance evaluation of various methods on BIQ2021 dataset: Boldface and italic text refer to the best and runner-up methods.

$$TSS = \sum_{i=1}^{n}(y_i - \bar{y})^2$$

where y_i is the actual value, $\bar{y}$ is the mean of the actual values, and n is the number of observations.

2. Compute the explained sum of squares (ESS): This represents the variance explained by the model.

$$ESS = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$$

where $\hat{y}_i$ is the predicted value.

3. The EVS is then calculated as the ratio of ESS to TSS:

$$\text{EVS} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

Where, y_{true} is the ground truth and y_{pred} is the predicted values.

Table 6 provides a summary of the EVS for various algorithms on the four datasets. Among the methods that were compared, the 'Proposed Method' shows itself to be the most effective algorithm with the best-explained Variance Score. This demonstrates the effectiveness of the 'Proposed Method' in explaining variance and outperforming other algorithms in this domain.

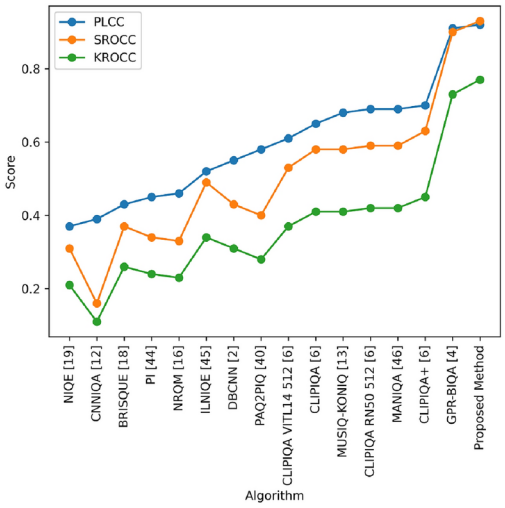
Quantile regression loss (QRL)
This section presents the results for QRL across different algorithms and datasets. QRL assesses the accuracy of predicted quantiles, offering insights into error distribution. In traditional regression tasks, the objective is to predict a point estimate (mean or median) of the target variable. However, in certain applications, comprehending prediction uncertainty is essential. Quantile loss offers a means to gauge this uncertainty by evaluating how effectively a model captures various quantiles of the target variable distribution.

Given a prediction $\hat{y}$ and the true target value y, the quantile loss at a specific quantile τ is defined as:

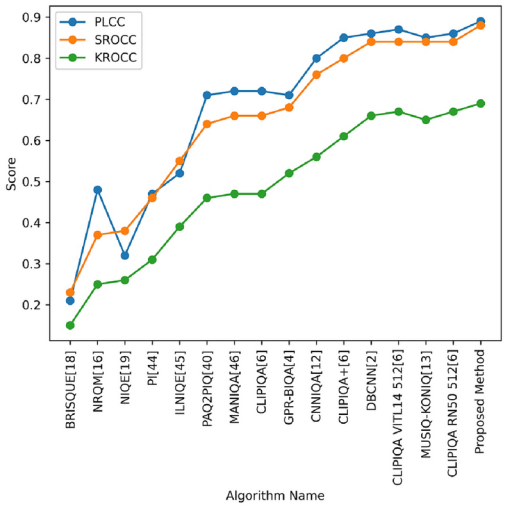
$$L_\tau(y, \hat{y}) = (\tau - I(y \leq \hat{y})) \cdot (\hat{y} - y)$$

Here:

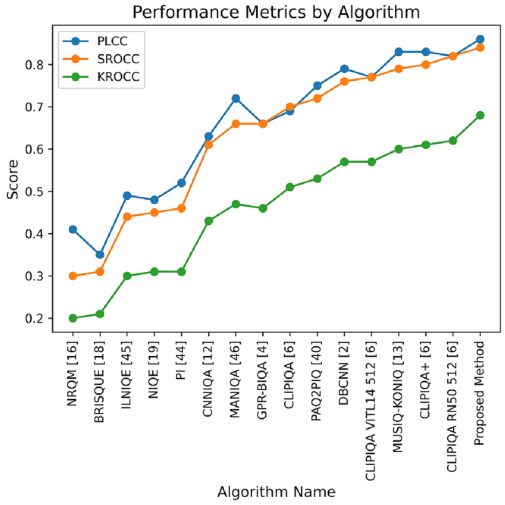
- τ is the quantile level (e.g., 0.1 for the 10th percentile, 0.5 for the median, 0.9 for the 90th percentile).



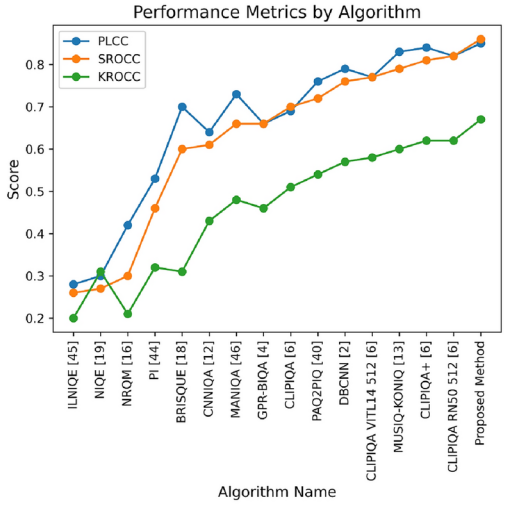
(a) Performance evaluation of various methods on TID2013 dataset



(b) Performance evaluation of various methods on KonIQ-10k dataset



(c) Performance evaluation of various methods on LiveC dataset



(d) Performance evaluation of various methods on BIQ2021 dataset

Fig. 6. Comparative Analysis of all algorithms with the proposed algorithm on various datasets

- $I(\cdot)$ is the indicator function (equals 1 if the condition is true, 0 otherwise).

The quantile loss penalizes the model more when it underestimates the true target value for lower quantiles ($y < \hat{y}$) and when it overestimates for higher quantiles ($y > \hat{y}$). It is asymmetric and provides a way to measure how well a model captures the tails or specific regions of the distribution. The quantile loss-based performance analysis on each of the datasets is reported in Figure7, 8, 9 and 10. Here, it can be noted that the proposed algorithm outperforms the other algorithms.

Residual analysis

A thorough residual analysis was carried out to gain a further understanding of the effectiveness of the proposed approach. Residuals, defined as the differences between true and predicted values, were examined at various quantiles to evaluate the model's behavior under varying conditions. In each dataset, residuals were calculated for three distinct quantiles (0.1, 0.5, and 0.9). These values were calculated using the QRL, which provides a thorough examination of algorithmic performance across various segments of the data distribution.

The resulting plots from the residual analysis, depicted in Figure 11, visually demonstrate how well each algorithm captures the intricacies of the data distribution. Each bar in the plot represents residuals for a specific combination of algorithm and quantile, with distinct colors distinguishing between quantiles. Interestingly, the proposed algorithm exhibits superior performance compared to others.

Algorithm	TID2013	KonIQ-10K	LiveC	BIQ2021
NIQE[58]	-6.30	-8.77	-3.34	-10.11
CNNIQA[18]	-5.57	-0.56	-1.52	-1.44
BRISQUE[59]	-4.41	-21.68	-7.16	-1.04
PI[60]	-3.94	-3.53	-2.70	-2.56
NRQM[61]	-3.73	-3.34	-4.95	-4.67
ILNIQE[33]	-2.70	-2.70	-3.16	-11.76
DBCNN[9]	-2.31	-0.35	-0.60	-0.60
PAQ2PIQ[50]	-1.97	-0.98	-0.78	-0.73
CLIPIQA_VITL14_512[6]	-1.69	-0.32	-0.69	-0.69
CLIPIQA[6]	-1.37	-0.93	-1.10	-1.10
MUSIQ-KONIQ[19]	-1.16	-0.38	-0.45	-0.45
CLIPIQA_RN50_512[6]	-1.10	-0.35	-0.49	-0.49
MANIQA[62]	-1.10	-0.93	-0.93	-0.88
CLIPIQA+[6]	-1.04	-0.38	-0.45	-0.42
GPR-BIQA[11]	-0.21	-0.98	-1.30	-1.30
Proposed Method	**-0.18**	**-0.29**	**-0.35**	**-0.38**

Table 6. EVS for various algorithms: Best score reported in boldface.

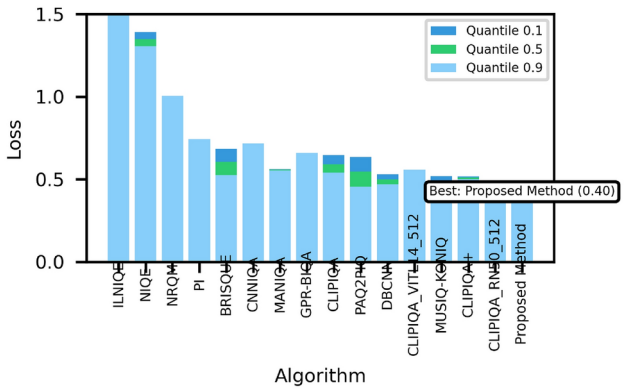


Fig. 7. QRL: Comparative analysis of various algorithms on BIQ2021 dataset

Qualitative analysis

The qualitative analysis of the proposed algorithm delves into the nuanced aspects of its performance beyond quantitative metrics. By examining visual outputs and subjective assessments, this analysis aims to provide a deeper understanding of how the algorithm handles various image distortions and artifacts. Through qualitative evaluation, is intended to uncover the algorithm's strengths, limitations, and overall effectiveness in accurately assessing image quality in real-world scenarios.

Regression analysis

In this evaluation phase, the relationship between the predicted and ground-truth scores is examined by fitting the best line to the data. The slope, intercept, and R-squared value of this line can be analyzed to gain insight into the algorithm's predictive accuracy and precision.

Figure 12 presents the results for the two largest datasets. The predicted line (best-fit line on scatter plot of prediction vs ground-truth data) closely aligns with the target line, indicating a high correlation between the predicted and ground-truth scores. This achievement fulfills the study's objective of designing an algorithm with a strong correlation with human judgment. The successful outcome highlights the algorithm's effectiveness in accurately predicting image quality, making it a valuable tool for blind image quality assessment tasks.

Distribution analysis

In the regression analysis, the distribution of the predicted variable holds significant importance as it should closely align with the distribution of the ground-truth. To address this concern, an analysis was conducted to examine the distributions of both variables. This analysis focused on the two largest authentically distorted datasets, mirroring the regression analysis approach. The results are depicted in Figure 13.

The analysis revealed that the predicted variable's density distribution closely matches that of the ground-truth variable, exhibiting similar mean, standard deviation, and type. This observation suggests that the

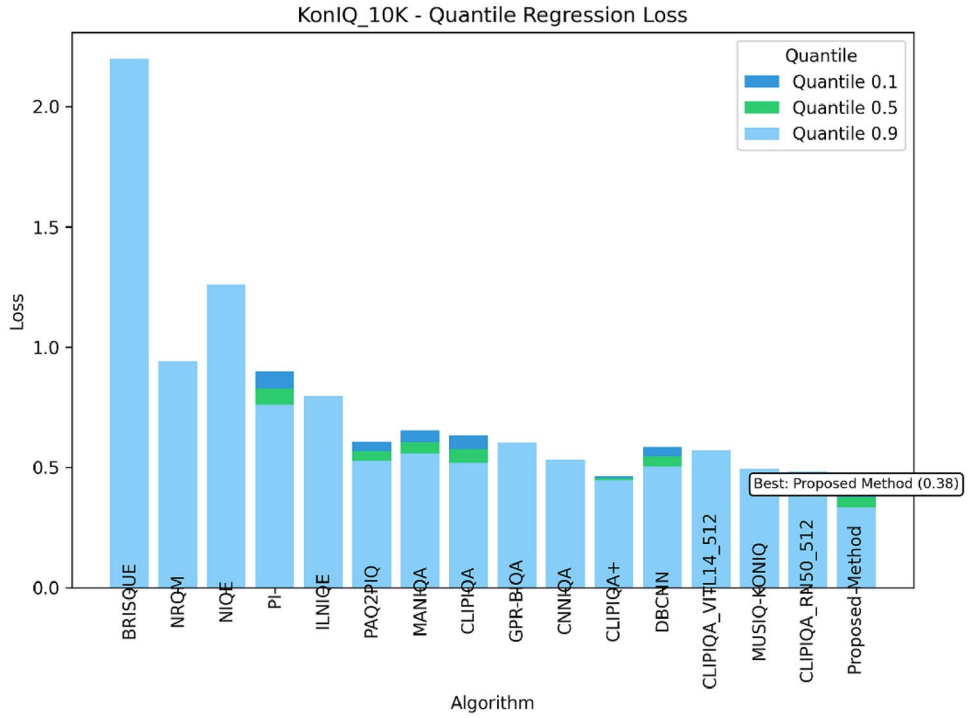


Fig. 8. QRL: Comparative analysis of various algorithms on Koniq-10k dataset

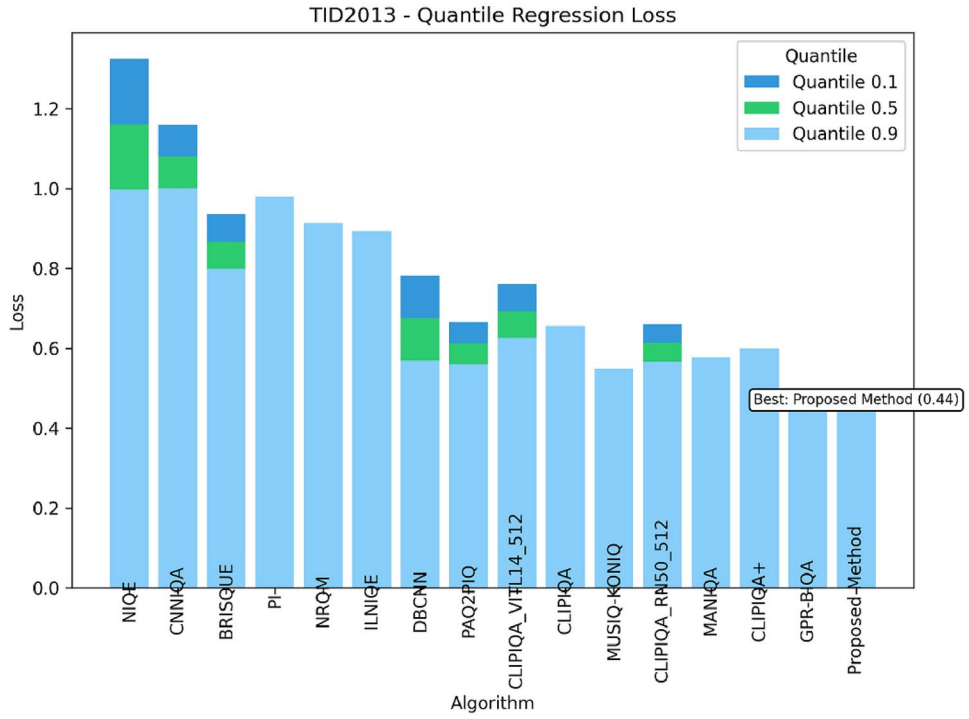


Fig. 9. QRL: Comparative analysis of various algorithms on TID2013 dataset TID2013 dataset

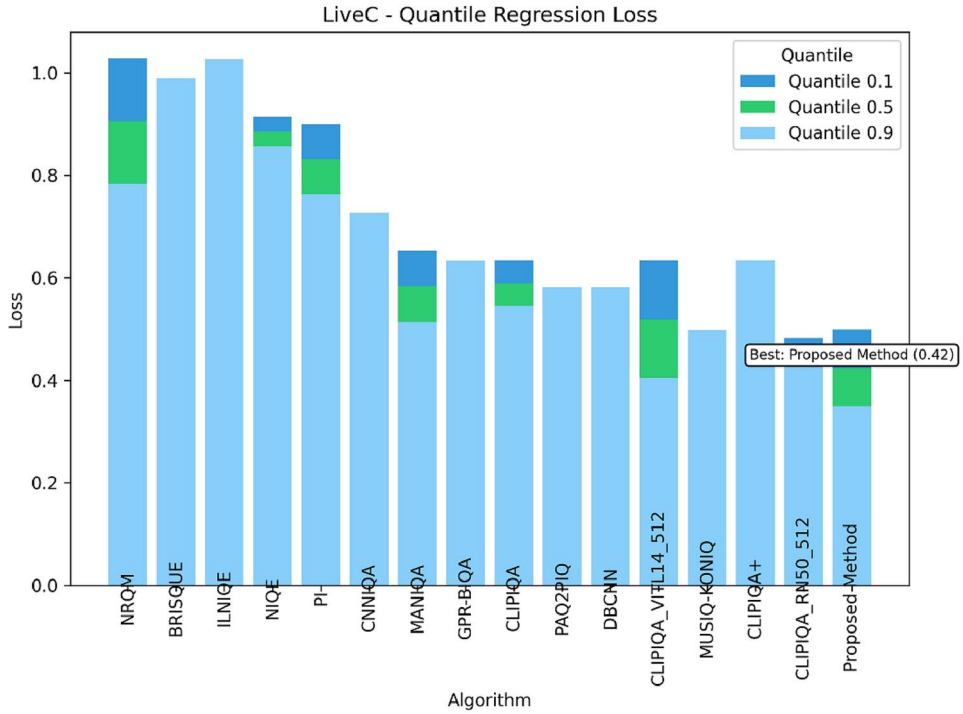


Fig. 10. QRL: Comparative analysis of various algorithms on LiveC dataset dataset

proposed algorithm effectively represents human perceptual behavior in the assessment of image quality. The striking resemblance in distributions validates the algorithm's capability to accurately predict image quality, thus enhancing its reliability for blind image quality assessment tasks.

Grad-CAM based visualization

To gain insights into the impact of attention and the effectiveness of the proposed architecture, Grad-CAM[63] was applied to a random set of images. Observations reveal that the proposed method puts greater weight on foreground information over the background. In Figure 14 (b) and (d), it is evident that the algorithm primarily focuses on information that directly affects perception, such as text, faces, and eyes. This behavior aligns with how humans perceive and understand distortions, as attention is naturally drawn to specific regions of interest.

The attention maps generated by the proposed algorithm are optimized, effectively highlighting crucial information for image quality assessment. The results demonstrate that the algorithm's attention mechanism is effective and generalizable, showcasing its ability to perform well on diverse image datasets. The efficiency and strong generalization capabilities further validate the robustness and reliability of the proposed approach. Consequently, the proposed algorithm's attention maps efficiently capture essential image details and features, leading to accurate quality assessments. These findings affirm the algorithm's practical utility and its capability to mimic human perception in blind image quality assessment tasks.

Ablation study

Ablation studies in deep neural networks provide valuable insights into the relationships between model parameters. Through the selective removal of specific parts of the network, a better understanding of their impact on performance and accuracy can be gained. In the context of the proposed algorithm, an ablation study is conducted, and the results are depicted in Figure 15. The experimentation involved the use of a train-test split of the KonIQ-10k dataset (one of the diverse and largest BIQA datasets).

The most important factors influencing the algorithm's performance can be found by meticulously investigating the consequences of eliminating particular parts of the model. This process helps to optimize the model architecture and fine-tune its parameters, resulting in improved accuracy and efficiency in blind image quality assessment.

The ablation study provides crucial information for refining the algorithm and enhancing its overall effectiveness, further validating its robustness in accurately predicting image quality. These findings contribute to a deeper comprehension of the proposed approach and reinforce its suitability for practical applications in image quality assessment tasks.

Comparison of chains with regression heads

The initial comparison between Chain-1 and Chain-2 inspired by (Resnet-50 and EfficentnetB7-inspired networks), each equipped with a regression head, revealed noteworthy insights. Both chains exhibited relatively high correlation coefficients, with Chain-1 achieving a PLCC of 0.83 and SROCC of 0.79, while Chain-2 closely

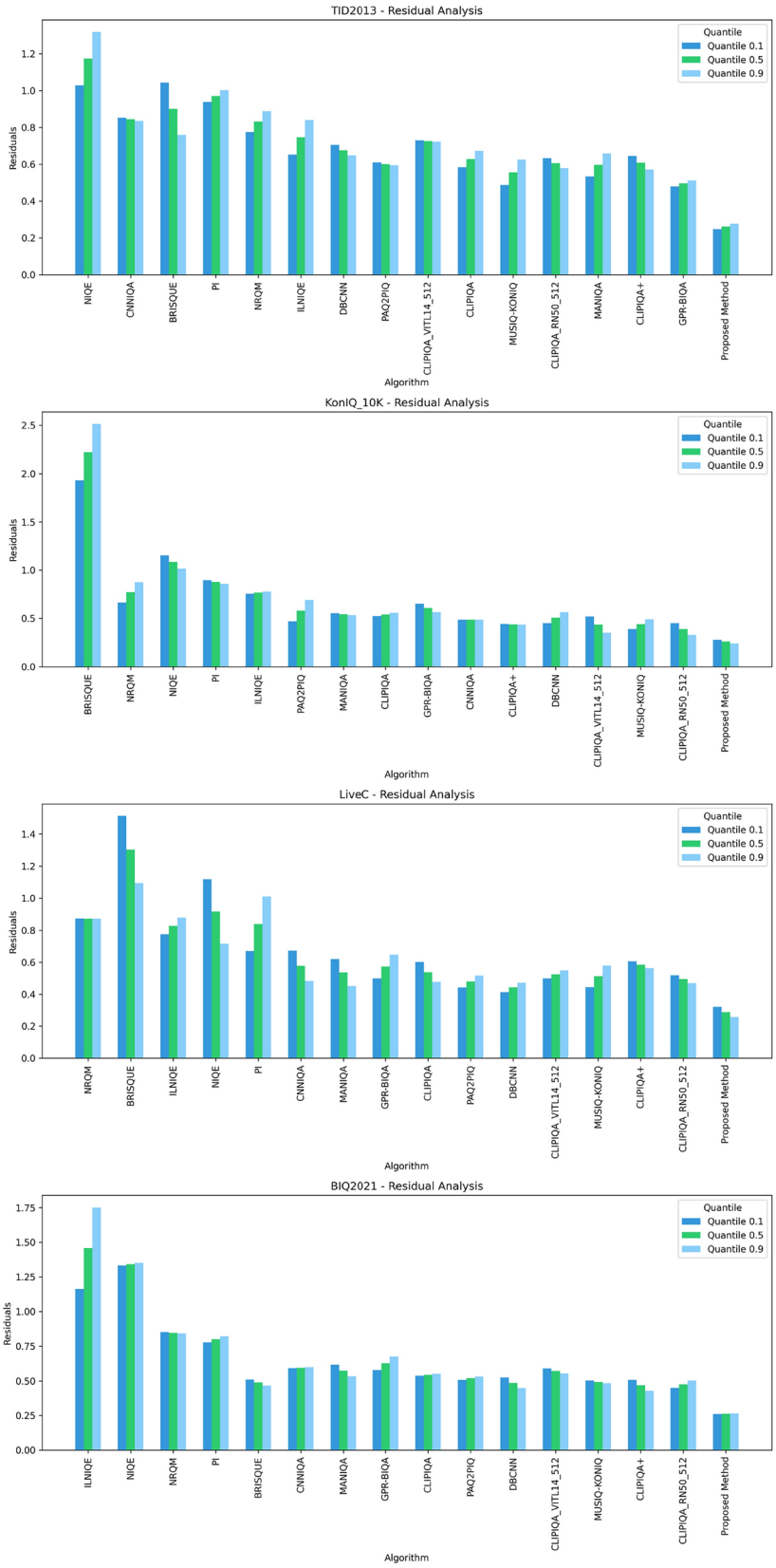


Fig. 11. Residual analysis for various algorithms at different quantiles.

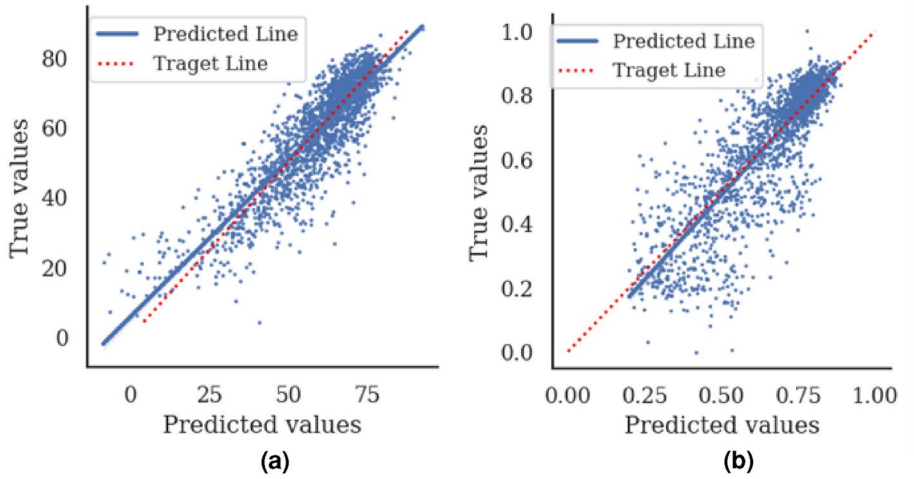


Fig. 12. Scatter-plot between ground-truth versus predicted values along with regression line for (a) KonIQ-10k (b) BIQ2021

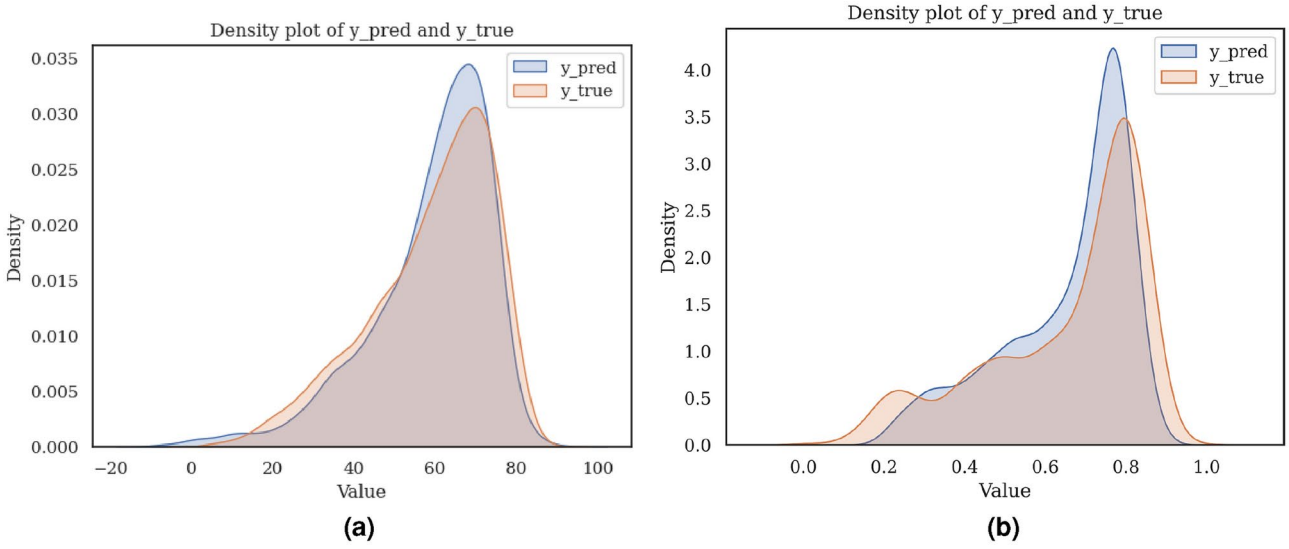


Fig. 13. Distribution plot between ground-truth versus predicted values for (**a**) KonIQ-10k (**b**) BIQA2021

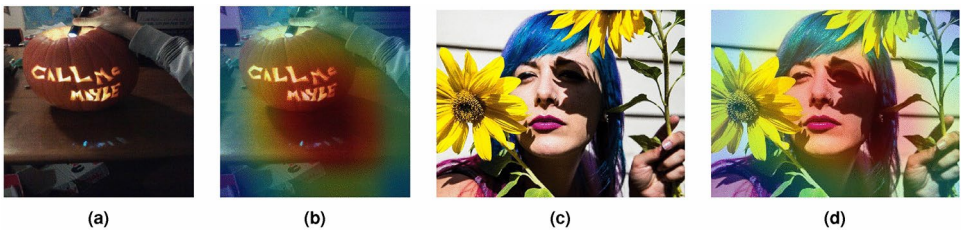


Fig. 14. Grad-CAM based heatmaps

followed with PLCC of 0.82 and SROCC of 0.78. This suggests that the regression head in both chains contributes significantly to the predictive capabilities.

Impact of feature fusion

Introducing feature fusion to Chain-1 and Chain-2 demonstrated a slight performance improvement. The combination of the two chains with feature fusion and regression heads resulted in increased PLCC (0.84) and

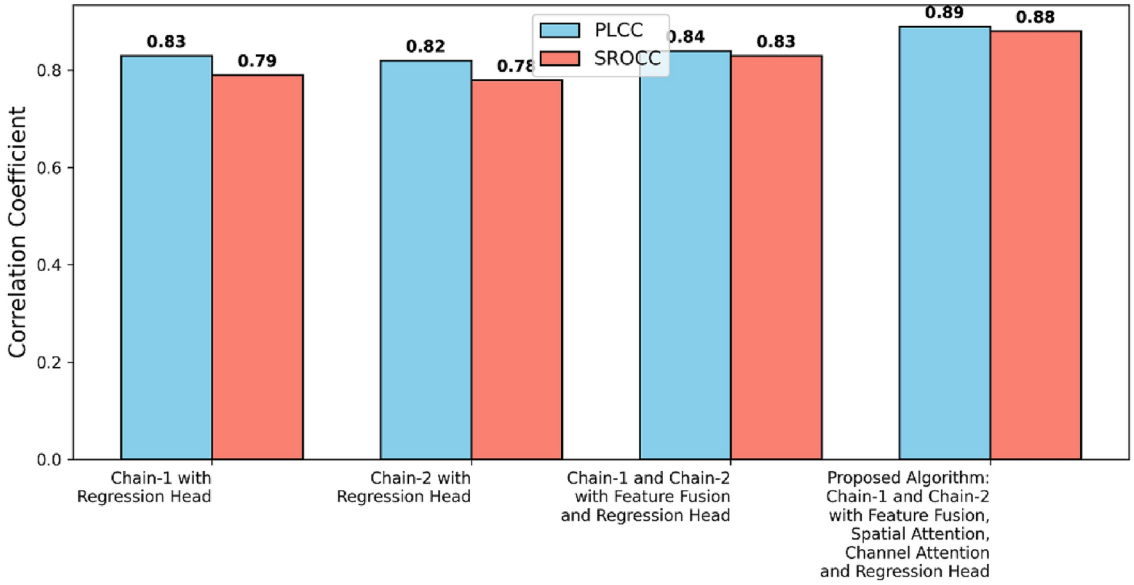


Fig. 15. Comparison of PLCC and SROCC for ablation study

SROCC (0.83). This indicates that combining features from both chains enhances the overall predictive capacity, supporting the effectiveness of feature fusion techniques.

Comprehensive approach: proposed Algorithm

The proposed algorithm, incorporating feature fusion, spatial attention, channel attention, and a regression head, exhibited the highest correlation coefficients among all configurations. The PLCC of 0.89 and SROCC of 0.88 signify a substantial enhancement in prediction accuracy compared to individual chains and the feature fusion scenario. The spatial and channel attention mechanisms, combined with feature fusion, demonstrate their efficacy in capturing intricate patterns within the data, contributing to the superior performance of the proposed algorithm.

Disscussion

The proposed model significantly advances image quality assessment (IQA) through a dual-backbone architecture that integrates ResNet and EfficientNet(ResNet for its robustness in handling complex textures and EfficientNet for its efficiency in feature extraction across multiple scales), optimizing feature extraction while reducing computational demands. Unlike self-supervised and continual learning approaches, which require extensive unlabeled datasets and complex training regimes, our model utilizes transfer learning to efficiently leverage pre-trained networks specifically for IQA tasks. While transformer-based models like the Swin Transformer excel in capturing intricate data relationships through elaborate self-attention mechanisms, they incur substantial computational costs. In contrast, our model employs streamlined spatial and channel attention mechanisms, refining feature representations without the overhead associated with multi-head attention layers. The use of simple concatenation for feature fusion preserves the richness of extracted features while minimizing overfitting risks. This fixed architecture ensures stability and adaptability, addressing the challenges of catastrophic forgetting found in continual learning frameworks. Ultimately, the proposed model strikes a harmonious balance between performance and efficiency, offering a practical alternative that enhances accuracy in IQA while remaining computationally feasible.

Conclusion

This study introduces a novel multistream fusion network with spatial and channel attention for blind image quality assessment. The proposed architecture undergoes comprehensive quantitative and qualitative evaluations and is compared against well-known state-of-the-art algorithms in the field. Through rigorous investigations, the proposed algorithm demonstrates superior performance, surpassing existing methods across renowned correlation coefficients on both large-scale authentic and synthetic distortion datasets. The efficacy of the spatial and channel attention mechanisms is further confirmed through attention visualization using Grad-CAM, revealing the algorithm's capability to focus on critical regions of the image and make decisions based on relevant information. The attention maps effectively highlight significant image features, contributing to the algorithm's impressive performance in quality assessment.

Significantly, the proposed algorithm exhibits a high correlation with human opinion, indicating its ability to accurately emulate human perceptual judgments. This characteristic positions it as a strong candidate for blind image quality assessment, with practical applications in various domains where reliable and objective image quality evaluation is essential. Resultantly, the proposed multistream fusion network with spatial and channel

attention represents a promising advancement in the field of blind image quality assessment, providing a robust and effective solution for real-world image analysis challenges.

Data availibility

The data used in this research is publically available for research and development purpose at the following links. **TID2013:** https://www.ponomarenko.info/tid2013.htm, **LiveC:** http://live.ece.utexas.edu/research/ChallengeDB/, **KonIQ-10K:** https://database.mmsp-kn.de/koniq-10k-database.html, **BIQ2021:** https://github.com/nisarahmedrana/BIQ2021.

Received: 23 April 2024; Accepted: 18 October 2024
Published online: 29 October 2024

References

1. Zhang, W. et al. Continual learning for blind image quality assessment. *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2022).
2. Ahmed, N., Asif, H. M. S., Saleem, G. & Younus, M. U. Image quality assessment for foliar disease identification (agropath). *J. Agric. Res* **59**, 177–186 (2021).
3. Ahmad, N. et al. Leaf image-based plant disease identification using color and texture features. *Wireless Personal Communications* **121**, 1139–1168 (2021).
4. Akhtar, M. Automated analysis of visual leaf shape features for plant classification. *Computers and Electronics in Agriculture* **157**, 270–280 (2019).
5. Nawaz, S., Calefati, A., Ahmed, N. & Gallo, I. Hand written characters recognition via deep metric learning. In *2018 13th IAPR International Workshop on Document Analysis Systems (DAS)*, 417–422 (IEEE, 2018).
6. Wang, J., Chan, K. C. & Loy, C. C. Exploring clip for assessing the look and feel of images. arXiv preprint arXiv:2207.12396 (2022).
7. Saleem, G., Bajwa, U. I. & Raza, R. H. Toward human activity recognition: a survey. *Neural Computing and Applications* **35**, 4145–4182 (2023).
8. Wang, Z., Bovik, A. C., Sheikh, H. R. & Simoncelli, E. P. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing* **13**, 600–612 (2004).
9. Zhang, W., Ma, K., Yan, J., Deng, D. & Wang, Z. Blind image quality assessment using a deep bilinear convolutional neural network. *IEEE Transactions on Circuits and Systems for Video Technology* **30**, 36–47 (2018).
10. Ahmed, N. & Asif, S. Biq 2021: a large-scale blind image quality assessment database. *Journal of Electronic Imaging* **31**, 053010–053010 (2022).
11. Khalid, H., Ali, M. & Ahmed, N. Gaussian process-based feature-enriched blind image quality assessment. *Journal of Visual Communication and Image Representation* **77**, 103092 (2021).
12. Aslam, M. A. & Ahmed, N. *& Saleem, G* (Visual representation learning for image quality assessment. IEEE Access, Vrl-iqa, 2023).
13. Zhai, G. & Min, X. Perceptual image quality assessment: a survey. *Science China Information Sciences* **63**, 1–52 (2020).
14. Aslam, M. A. et al. Tqp: An efficient video quality assessment framework for adaptive bitrate video streaming. *IEEE Access* **12**, 88264–88278. https://doi.org/10.1109/ACCESS.2024.3418375 (2024).
15. Ahmed, N., Shahzad Asif, H., Bhatti, A. R. & Khan, A. Deep ensembling for perceptual image quality assessment. *Soft Computing* **26**, 7601–7622 (2022).
16. Sheikh, H. R. Live image quality assessment database. http://live.ece.utexas.edu/research/quality (2003).
17. Hosu, V., Lin, H., Sziranyi, T. & Saupe, D. Koniq-10k: An ecologically valid database for deep learning of blind image quality assessment. *IEEE Transactions on Image Processing* **29**, 4041–4056 (2020).
18. Kang, L., Ye, P., Li, Y. & Doermann, D. Convolutional neural networks for no-reference image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 1733–1740 (2014).
19. Ke, J., Wang, Q., Wang, Y., Milanfar, P. & Yang, F. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 5148–5157 (2021).
20. Larson, E. C. & Chandler, D. M. Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging* **19**, 011006–011006–21 (2010).
21. Ma, K. et al. End-to-end blind image quality assessment using deep neural networks. *IEEE Transactions on Image Processing* **27**, 1202–1213 (2017).
22. Ahmed, N., Asif, H. M. S. & Khalid, H. Non-reference quality monitoring of digital images using gradient statistics and feedforward neural networks. arXiv preprint arXiv:2112.13893 (2021).
23. Saleem, G., Bajwa, U. I., Raza, R. H. & Zhang, F. Edge-enhanced tempofusenet: A two-stream framework for intelligent multiclass video anomaly recognition in 5g and iot environments. *Future Internet* **16**, 83 (2024).
24. Saleem, G. et al. Efficient anomaly recognition using surveillance videos. *PeerJ Computer Science* **8**, e1117 (2022).
25. Zaman, M. I., Bajwa, U. I., Saleem, G. & Raza, R. H. A robust deep networks based multi-object multi-camera tracking system for city scale traffic. *Multimedia Tools and Applications* **83**, 17163–17181 (2024).
26. Iqbal, S., Naveed, K., Naqvi, S. S., Naveed, A. & Khan, T. M. Robust retinal blood vessel segmentation using a patch-based statistical adaptive multi-scale line detector. *Digital Signal Processing* **139**, 104075 (2023).
27. Naveed, A., Naqvi, S. S., Khan, T. M. & Razzak, I. Pca: Progressive class-wise attention for skin lesions diagnosis. *Engineering Applications of Artificial Intelligence* **127**, 107417 (2024).
28. Saleem, G., Bajwa, U. I. & Raza, R. H. Surveilia: Anomaly identification using temporally localized surveillance videos. *Available at SSRN 4308311* (2024).
29. Moorthy, A. K. & Bovik, A. C. Blind image quality assessment: From natural scene statistics to perceptual quality. *IEEE transactions on Image Processing* **20**, 3350–3364 (2011).
30. Saad, M. A., Bovik, A. C. & Charrier, C. Blind image quality assessment: A natural scene statistics approach in the dct domain. *IEEE transactions on Image Processing* **21**, 3339–3352 (2012).
31. Xue, W., Zhang, L. & Mou, X. Learning without human scores for blind image quality assessment. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 995–1002 (2013).
32. Xue, W., Mou, X., Zhang, L., Bovik, A. C. & Feng, X. Blind image quality assessment using joint statistics of gradient magnitude and laplacian features. *IEEE Transactions on Image Processing* **23**, 4850–4862 (2014).
33. Zhang, L., Zhang, L. & Bovik, A. C. A feature-enriched completely blind image quality evaluator. *IEEE Transactions on Image Processing* **24**, 2579–2591 (2015).
34. Fang, Y. et al. No-reference quality assessment of contrast-distorted images based on natural scene statistics. *IEEE Signal Processing Letters* **22**, 838–842 (2014).
35. Gu, K., Zhai, G., Yang, X., Zhang, W. & Liu, M. Subjective and objective quality assessment for images with contrast change. In *2013 IEEE International Conference on Image Processing*, 383–387 (IEEE, 2013).

36. Xu, J. et al. Blind image quality assessment based on high order statistics aggregation. *IEEE Transactions on Image Processing* **25**, 4444–4457 (2016).
37. Ghadiyaram, D. & Bovik, A. C. Perceptual quality prediction on authentically distorted images using a bag of features approach. *Journal of vision* **17**, 32–32 (2017).
38. Ghadiyaram, D. & Bovik, A. C. Live in the wild image quality challenge database. Online: http://live.ece.utexas.edu/research/ChallengeDB/index.html[Mar, 2017] (2015).
39. Kundu, D., Ghadiyaram, D., Bovik, A. C. & Evans, B. L. No-reference quality assessment of tone-mapped hdr pictures. *IEEE Transactions on Image Processing* **26**, 2957–2971 (2017).
40. Sadiq, A., Nizami, I. F., Anwar, S. M. & Majid, M. Blind image quality assessment using natural scene statistics of stationary wavelet transform. *Optik* **205**, 164189 (2020).
41. Ahmed, N., Asif, H. M. S. & Khalid, H. Piqi: perceptual image quality index based on an ensemble of gaussian process regression. *Multimedia Tools and Applications* **80**, 15677–15700 (2021).
42. Alotaibi, T., Khan, I. R. & Bourennani, F. Quality assessment of tone-mapped images using fundamental color and structural features. *IEEE Transactions on Multimedia* **26**, 1244–1254. https://doi.org/10.1109/TMM.2023.3278989 (2024).
43. Gu, K., Zhai, G., Yang, X. & Zhang, W. Deep learning network for blind image quality assessment. In *2014 IEEE International Conference on Image Processing (ICIP)*, 511–515 (IEEE, 2014).
44. Fu, J., Wang, H. & Zuo, L. Blind image quality assessment for multiply distorted images via convolutional neural networks. In *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 1075–1079 (IEEE, 2016).
45. Bianco, S., Celona, L., Napoletano, P. & Schettini, R. On the use of deep learning for blind image quality assessment. *Signal, Image and Video Processing* **12**, 355–362 (2018).
46. Larson, E. C. & Chandler, D. Categorical image quality (csiq) database (2010).
47. Ponomarenko, N. et al. Tid 2008-a database for evaluation of full-reference visual quality assessment metrics. *Advances of modern radioelectronics* **10**, 30–45 (2009).
48. Ponomarenko, N. *et al.* Color image database tid2013: Peculiarities and preliminary results. In *European workshop on visual information processing (EUVIP)*, 106–111 (IEEE, 2013).
49. Ahmed, N. & Asif, H. M. S. Ensembling convolutional neural networks for perceptual image quality assessment. In *2019 13th International Conference on Mathematics, Actuarial Science, Computer Science and Statistics (MACS)*, 1–5 (IEEE, 2019).
50. Ying, Z. *et al.* From patches to pictures (paq-2-piq): Mapping the perceptual space of picture quality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 3575–3585 (2020).
51. Ahmed, N. & Asif, H. M. S. Perceptual quality assessment of digital images using deep features. *Computing & Informatics* **39** (2020).
52. Varga, D. No-reference image quality assessment with convolutional neural networks and decision fusion. *Applied Sciences* **12**, https://doi.org/10.3390/app12010101 (2022).
53. Sang, Q., Shu, Z., Liu, L., Hu, C. & Wu, Q. Image quality assessment based on self-supervised learning and knowledge distillation. *Journal of Visual Communication and Image Representation* **90**, 103708 (2023).
54. Li, X. & He, S. Blind image quality evaluation method based on cyclic generative adversarial network. *IEEE Access* (2024).
55. Yang, Y., Lei, Z. & Li, C. No-reference image quality assessment combining swin-transformer and natural scene statistics. *Sensors* **24**, https://doi.org/10.3390/s24165221 (2024).
56. Zamir, S. W. et al. Learning enriched features for fast image restoration and enhancement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*[SPACE]https://doi.org/10.1109/TPAMI.2022.3167175 (2022).
57. Ghadiyaram, D. & Bovik, A. C. Massive online crowdsourced study of subjective and objective picture quality. *IEEE Transactions on Image Processing* **25**, 372–387 (2015).
58. Mittal, A., Soundararajan, R. & Bovik, A. C. Making a "completely blind" image quality analyzer. *IEEE Signal processing letters* **20**, 209–212 (2012).
59. Mittal, A., Moorthy, A. K. & Bovik, A. C. No-reference image quality assessment in the spatial domain. *IEEE Transactions on image processing* **21**, 4695–4708 (2012).
60. Blau, Y., Mechrez, R., Timofte, R., Michaeli, T. & Zelnik-Manor, L. The 2018 pirm challenge on perceptual image super-resolution. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, 0–0 (2018).
61. Ma, C., Yang, C.-Y., Yang, X. & Yang, M.-H. Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding* **158**, 1–16 (2017).
62. Yang, S. *et al.* 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 (2022).
63. Selvaraju, R. R. *et al.* Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626 (2017).

Acknowledgements

All authors thank the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun China and the School of Information Engineering, Xi'an Eurasia University, Xi'an, Shaanxi, China, for their Financial Support and Funding.

Author contributions

M.A.A., H.K., and N.A. conceived the research idea. N.A. and H.K. developed the methodology. M.A.A., X.W., Z.S., and X.L. conducted the experiments and analyzed the data. M.A.A., H.K., and N.A. interpreted the results. M.A.A., N.A. and H.K. drafted the manuscript. M.A.A., H.K., N.A., X.W., Z.S., X.L., and Y.X. reviewed and edited the manuscript. All authors approved the final version of the manuscript for submission.

Funding

The funding is provided by the Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China, Grant Number 2024 PVB0036 under the PIFI Visiting Scientists Program. All authors also thank the School of Information Engineering, Xi'an Eurasia University, Xi'an, Shaanxi, China, for their partial financial support.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to M.A.A.