Hardware Implementation of a Deep Learning-based Model for Image Quality Assessment

Yahia Said
Department of Electrical Engineering, College of Engineering, Northern Border University, Saudi Arabia
yahia.said@nbu.edu.sa (corresponding author)

Yazan A. Alsariera
Department of Computer Science, College of Science, Northern Border University, Saudi Arabia
yazan.sadeq@nbu.edu.sa

Received: 4 March 2024 | Revised: 13 March 2024 | Accepted: 14 March 2024

 DOI: https://doi.org/10.48084/etasr.7194

ABSTRACT

Image quality assessment is very important for accurate analysis and better interpretation. In reality, environmental effects and device limitations may degrade image quality. Recently, many image quality assessment algorithms have been proposed. However, these algorithms require high computation overhead, making them unsuitable for mobile devices, such as smartphones and smart cameras. This paper presents a hardware implementation of an image quality assessment algorithm based on a Lightweight Convolutional Neural Network (LCNN) model. Many advances have been made in the construction of high-accuracy LCNN models. The current study used EfficientNet V2. The model achieved state-of-the-art image classification performance on many famous benchmark datasets while having a smaller size than other models with the same performance. The model was utilized to learn human visual behavior through understanding dataset information without prior knowledge of target visual behavior. The proposed model was implemented employing a Field Programmable Gate Array (FPGA) for possible integration into mobile devices. The Xilinx ZCU 102 board was implemented to evaluate the proposed model. The results confirmed the latter's efficiency in image quality assessment compared to existing models.

Keywords-deep learning; artificial intelligence; embedded systems; FPGA; efficientNet v2; image quality assessment

I. INTRODUCTION

The need for a reliable and effective Image Quality Assessment (IQA) has increased significantly in recent years [1]. IQA has become a progressively attractive subject, with a huge number of new algorithms being developed each year. Based on the visual cues offered by the original reference image, IQA may be split into three categories: Full Reference-IQA (FR-IQA), Reduced Reference-IQA (RR-IQA), and No Reference-IQA (NR-IQA). FR-IQA denotes the availability of a deformed image and the original one, NR-IQA uses only distorted images, and RR-IQA utilizes a portion of the original picture's visual clues or certain attributes on the reference image. In most cases, the availability of the reference image is limited in practical situations. Due to the constantly changing picture content, in addition to the absence of a visual reference, NR-IQA is considered a highly demanding study case. NR-IQA may be classified into two main categories: distortion-based and general-purpose methods. A distortion-based method is adopted to assess the quality associated with a specific type of distortion, such as blocking or blurring. As it is not always possible to define the type of distortion, distortion-based techniques have a limited use in real-world circumstances.

Despite several efforts to build a viable NR-IQA algorithm for real-world applications, the performances achieved are below the expected. Many IQA techniques employ deep learning to build a correlation between the extracted quality attributes of the picture and the objective IQA score. These algorithms typically depend on features that are designed manually to be sensitive to image quality. Deep learning performs well in a wide variety of applications, including computer vision. Deep learning techniques require extensive computational resources to accomplish high performance. However, many models have been proposed to reduce computation complexity without degrading accuracy. The ever-increasing desire for solutions that are both effective and scalable in a variety of applications is the driving force behind the usage of lightweight deep learning models for IQA.

Traditional deep learning models, particularly those deployed for image processing tasks, frequently require a substantial amount of computing resources and memory. As a result, these models are not ideal for settings with limited resources, such as mobile devices or embedded systems. Lightweight models provide a compromise by delivering satisfactory performance while also reducing the amount of computing complexity required, making them suitable for low-

power devices. Lightweight models offer faster inference times, enabling quick decision-making and reaction in applications where image quality evaluation has to be performed in real-time. Some examples of these applications entail video streaming, surveillance systems, and augmented reality applications. Lightweight models are easier to implement and deploy across a wide variety of platforms and devices. It is possible to incorporate them into cloud-based systems, Internet of Things (IoT) devices, or edge computing devices without large infrastructure modifications. Lightweight models have fewer memory footprints and reduced energy use than their heavier counterparts. This is of utmost importance in mobile devices powered by batteries. It is common for lightweight models to generalize well across a variety of datasets and circumstances, which enables them to be acceptable for a wide range of quality assurance activities without compromising performance. Lightweight models can be trained and fine-tuned employing fewer datasets, making them accessible to a wider audience of researchers and developers who may not have access to large-scale datasets or high-performance computer resources. Therefore, lightweight deep learning models for IQA can achieve a balance between computational efficiency, performance, and scalability, making them applicable for a wide variety of real-world applications.

This study used the EfficientNet v2 model [2] to extract features from the proposed IQA algorithm. The proposed approach utilized three neural networks, following the human visual perception system that operates on a hierarchical basis. Extracting edge local information and substrate texture data was performed employing EfficientNet v2-S. Meanwhile, global semantic high-level features were extracted putting into service EfficientNet v2-M. The proposed model incorporates several novelties compared to other approaches currently used in the field of IQA. The proposed model is trained to learn human visual behavior by evaluating information from datasets without depending on prior knowledge of the target visual behavior. Due to its capacity for self-learning, the model can adapt and generalize effectively across a wide variety of IQA tasks and hypothetical situations. This study focused on the implementation of the IQA algorithm directly on hardware, deploying a Field-Programmable Gate Array (FPGA). This in contradicts with the usual software-based approaches, which sometimes include a significant amount of processing overhead. The particular approach provides benefits in terms of speed, energy efficiency, and the ability to perform real-time processing. This model engaged LCNNs, which are meant to have high accuracy while preserving minimal computational complexity and memory footprint. The EfficientNet v2 model, which is well-known for its exceptional performance in picture classification tasks while having a smaller model size, can guarantee the effective usage of computing resources without sacrificing accuracy. Therefore, this method is advisable for mobile devices, such as smartphones and smart cameras. The implementation of the IQA algorithm on FPGA hardware can facilitate the deployment of the model in real-world mobile applications that often have limited processing resources. The main contributions of this work are the following:

- Develop a deep CNN structure that can extract natural scene statistical data that are incredibly suggestive of human visual perception and cognition to perform an NR-IQA task which is straightforward but effective. As a result, both the spatial and spectral domains of image representation are improved.
- Present a distinctive IQA FPGA implementation. The hardware has been optimized for the pipeline throughout. Simultaneously, a small quantity of hardware is introduced in place of the usual large-capacity storage. In addition to improving the working rate and reducing the circuit scale, this arrangement can save space.

II. RELATED WORKS

Several important algorithms have been proposed for FR-IQA. PSNR [3] is often used to measure the quality of compressed pictures and is regarded as an IQA criterion in signal processing-related research disciplines. PSNR has low computing complexity and can be implemented quickly but cannot encode several critical features of the Human Vision System (HVS) [4]. A novel IQA technique called SSIM [5] has been proposed to improve HVS, while MS-SSIM [6] and IW-SSIM [7] have been recommended as IQA measures. The information from photos acquired under various situations and resolutions is combined and seamlessly integrated into the IQA procedure. The VIF approach [8] utilizes a source model, a distortion model, and an HVS model to leverage GSMs to describe realistic images in the wavelet domain.

Mechanisms for analyzing image distortion, such as MAD [10], were developed with the assumption that the HVS employs a variety of mechanisms to assess picture quality. Analyzing local brightness, contrast masking, and changes in the local statistical properties of the spatial frequency components are among these strategies [9]. According to the FSIM method [11], the visual system is highly dependent on gradient and phase consistency to determine image quality. These two elements are believed to be the fundamental ways in which HVS interprets images [17]. The FSIMc technique was established by adjusting the average depending on phase consistency data, including color attributes. In contrast, VSI [12] deploys a saliency map in place of FSIMc's phase consistency feature to enhance the results while retaining the color information and its gradient from FSIMc. GMSD [13] puts into service gradient as the key feature for assessing image quality and uses standard deviation pooling instead of the classic mean pooling. However, in real-world circumstances, the reference image may not be available and only parts of the visual cues or implicit visual aspects may be accessible. In such instances, the RR-IQA algorithm can provide a feasible solution. In [14], an NR-IQA method was presented utilizing grouped transformations.

III. PROPOSED ARCHITECTURE

A hybrid approach enables the extraction of both local and non-local characteristics from the image. Figure 1 depicts the main architecture of the proposed model. At first, a hierarchical feature extraction module is used to extract features from the input images. The three different feature extraction modules extract multilevel features at different stages. A texture extraction module, based on EfficientNet v2-S, extracts

ground-level texture features. Low-level edge features are extracted applying an edge extraction module based on EfficientNet v2-S. EfficientNet v2-M can extract sophisticated semantic features. Subsequently, an attention mechanism fuses the retrieved features.

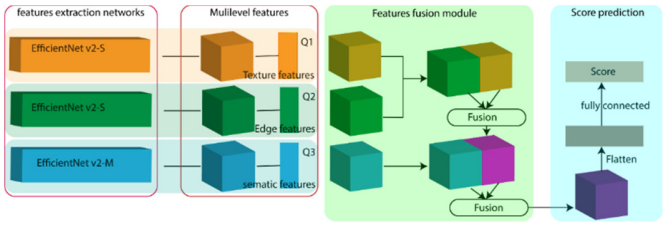

Fig. 1. Proposed model for NR-IQA.

The texture module provides low-level features, the edge extraction module captures local features, and the semantic features extraction module acquires high-level global semantic information. The dimensions of the feature maps, using the EfficientNet v2-S architecture in the attentional module, are 256×56×56. After undergoing the Conv 2d (512, 512, 1) and Conv 2d (256, 512, 1) operations, the dimensions of the texture feature map are altered to 512×7×7. The edge extraction network provides 1×224×224 feature maps. The proposed modification results in feature maps appear with a size of 640×7×7. Then, an additional convolution layer Conv 2d (2048, 512, 1,) is applied to generate 512×7×7 feature maps.

The EfficientNet v2-S presents seven stages after removing the classification head composed of convolution, pooling, and fully connected layers. The first stage is a 3×3 convolution layer, the second, third, and fourth stages are fused MBConv layers with a 3×3 kernel, and the last three stages are based on MBConv layers with a 3×3 kernel. To fix the number of channels feature maps size, a convolution layer followed by a deconvolution layer were applied to the last five stages from the top of the network. Figure 2 illustrates the proposed edge extraction network. To extract the global semantic information, the classification head of the EfficientNet v2-M was replaced with a convolution layer Conv 2d (2048, 512, 1) which generates feature maps with 512 channels.

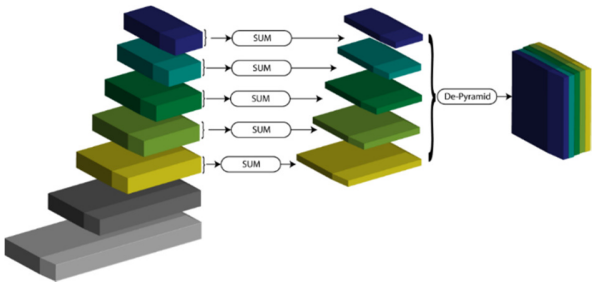

Fig. 2. Architecture of the proposed edge features extraction network.

Since the proposed three networks provide features with complex correlation, the fusion technique is very important for achieving good results. Applying a simple fusion technique, such as concatenation or addition, will certainly degrade the ability to consider multilevel features for image contain expression. To take advantage of the extracted features, an attention mechanism was adopted as a fusion technique. Three identical attention blocks make up the suggested fusion method. The architecture of the fusion approach, which is based on the attention mechanism, is shown in Figure 3.

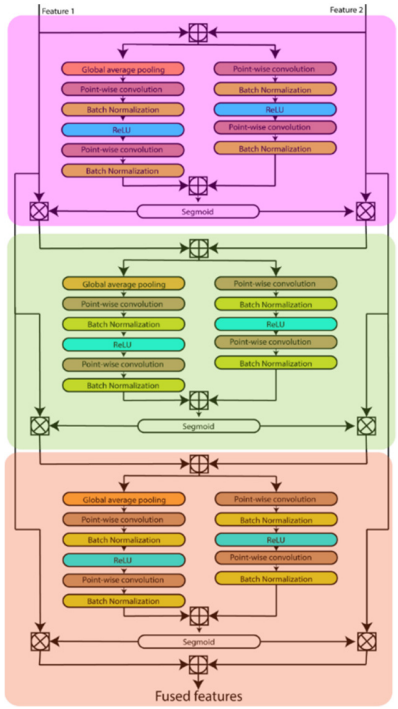

Fig. 3. Proposed fusion module.

Considering two sets of features with a size $H \times W$ and C channels $X, Y \in \mathbb{R}^{C \times H \times W}$, the fusion result Z based on the proposed attention mechanism is given by:

$$Z_1 = w_1(X \oplus Y) \otimes X + (1 - w_1(X \oplus Y) \otimes Y)$$
$$Z_2 = w_2(Z_1) \otimes X + (1 - w_1(Z_1) \otimes Y)$$
$$Z = w_3(Z_2) \otimes X + (1 - w_1(Z_2) \otimes Y \quad (1)$$

where Z represents fusion output, Z_1 is the result of the first attention block, Z_2 is the fusion result of the second attention block, $\otimes$ is element-wise multiplication, and $\oplus$ is broadcasting addition. w_1, w_2, and w_3 are the corresponding weight matrices for each attention block, which can be computed by:

$$w(M) = sig(l(M) \oplus g(M)) \quad (2)$$

where $sig(.)$is the sigmoid function, $M \in \mathbb{R}^{C \times H \times W}$ is the sum of x and Y, $g(M)$ is the context of the global feature, and $l(M)$ is the context of the local feature, which are computed as:

$$l(M) = b(pwconv2(r(b(pwconv1(M))))) \quad (3)$$

$$g(M) = b(D_i(r(b(D_r(g(M)))))) \quad (4)$$

where $pwconv$ are pointwise convolution layers with kernel sizes of $C/t \times C \times 1 \times 1$ for $pwconv1$ and $C \times C/t \times 1 \times 1$ for $pwconv2$, t is the reduction rate of the channel, r

refers to the Rectified Linear Unit (ReLU) function, b refers to the batch normalization, D_i is channel increasing layer, D_r is the channel reduction layer, and $g(M)$ is the global average pooling layer. To create the best match between the distorted picture and the score, it is necessary to minimize the difference between the anticipated score determined by objective criteria and the value given based on subjective judgment. The Huber loss was adopted to increase the network's robustness. The proposed loss function for NR-IQA can be calculated by:

$$L(s) = a^2 \left(\sqrt{1 + \left(\frac{s-m}{a} \right)^2} - 1 \right) \quad (5)$$

where m is the target score, s is the predicted score, and a is a hyperparameter. The extracted features are considered in loss optimization by measuring the difference between the score of each feature and the mean score. Three scores ($s_1, s_2, and\ s_3$) are associated with the hierarchical features and the final score is s_f. The main loss function is given by:

$$L = L_s + L_f = \sum_{i=1}^{3} b_i L(s_i) + L(s_f) \quad (6)$$

where L_f represents the loss function that measures the discrepancy between the predicted and the target score, L_s is the aggregation of different scores among the extracted features, and b_i is a relative weight of the score associated with each feature which is calculated by:

$$b_i = (sig(s_i) - 1/2)^2 \quad (7)$$

where $sig(s_i)$is the sigmoid function given by:

$$sig(s_i) = \frac{1}{1+\exp(-(s_i - m))} \quad (8)$$

Considering the propriety of the sigmoid function, if there is a small variation between s_i and m, then the corresponding value of b_i is small. The relative weights are presented as:

$$(b_1, b_2, b_3)^* = \arg\min(L(s;\ b_1, b_2, b_3)) \quad (9)$$

Optimizing the main loss function results in finding the optimal parameters for the proposed model. Algorithm 1 presents the main flow of the proposed model.

```
Algorithm 1
Input:
  Images = {i_1, i_2, ..., i_n}
  Labels = {m_1, m_2, ..., m_n}
Output:
  Score s
N: Number of train iterations
For n =1 to N do
  x_n^k = CNN_k(x_{n-1}^k);
  Z_n = Z(x_n^1, x_n^2, x_n^3);
  s_n = FC(Z_n);
  Loss = L(s, s_1, s_2, s_3);
End
x^k = CNN_k(image);
Z = Z(x^1, x^2, x^3);
s = FC(Z);
```

In the training at iteration n, each network generates a result x_n^i which is used as input to the fusion module to generate Z_n. The output of the fusion module is then passed through the fully connected layers to predict the final score s_n. At each iteration, the parameters are optimized continuously by jointly considering the scores of each network of the model and the final score. In the testing process, the input image is passed through the model using the optimized parameters to predict the quality score.

IV. EXPERIMENTS AND RESULTS

A. Data and Evaluation Metrics

The proposed model was evaluated on seven different public datasets to prove its efficiency: LIVE [15], CSIQ, TID2013 [16], KADID-10K [18], CLIVE [19], KonIQ-10k, and LIVE-FB [20]. The first four databases' image distortion is artificial synthesis, whereas the latter three databases' image distortion is natural scene distortion. Table I contains information on the seven datasets. The LIVE dataset includes 799 distorted images affected by 5 different distortion types. Distorted images have many resolutions such as 640×512, 768×512, and 480×720. The KADID dataset has a total of 10125 distorted images, generated by applying 25 distortion types, each one with five levels. The generated images have a resolution of 500×500. The TID2013 dataset contains 3,000 images, obtained by applying 24 different distortion types, each with 5 levels, on 25 original images. The CSIQ dataset provides a total of 866 distorted images based on the influence of 6 types of distortion on 30 original images. The CLIVE dataset contains 1,162 naturally distorted images with a resolution of 500×500. The KonIQ dataset was naturally collected and has 10,073 images with a resolution of 512×384. The LIVEFB is the largest NR-IQA dataset with 39,810 images collected from natural scenes with real distortion.

TABLE I. STATISTICS ON THE DATASETS USED

Dataset	Distorted images	Number of Distortion	Distortion
LIVE	799	5	Synthetic
CSIQ	866	6	Synthetic
TID2013	3000	24/ 5 levels	Synthetic
KADID	10125	24/ 5 levels	Synthetic
CLIVE	1162	-	Authentic
KonIQ	10073	-	Authentic
LIVEFB	39810	-	Authentic

For comparison purposes, two standard evaluation metrics were used. The first is the Pearson Linear Correlation Coefficient (PLCC), which computes the correlation between two scores. A high PLCC value indicates a good correlation between the subjective and objective scores. PLCC is given by:

$$PLCC = \frac{\sum_{i=1}^{N}(m_i - \bar{m})(s_i - \bar{s})}{\sqrt{\sum_{i=1}^{N}(m_i - \bar{m})^2}\sqrt{\sum_{i=1}^{N}(s_i - \bar{s})^2}} \quad (10)$$

where m_i is the mean opinion score, s_i is the objective score for an input image i, $\bar{s}$ is the objective score, and $\bar{m}$ represents the mean value of the subjective score. The second is the Spearman Rank Order Correlation Coefficient (SROCC), which calculates the correlation between the objective and subjective scores. A high SROCC value indicates a good correlation between subjective and objective scores. SROCC can be calculated by:

$$SROCC = 1 - \frac{6\sum_{i=1}^{N} d_i^2}{N(N^2-1)} \quad (11)$$

where d_i represents the discrepancy between the ranking of image i based on objective and subjective quality scores.

B. Implementation Details

The proposed model was developed based on the Vitis AI framework with the support of the TensorFlow deep learning library. The model was trained on a Nvidia GTX 960 GPU and tested on the Xilinx ZCU 102 FPGA. The Adam optimizer was used as a training algorithm with a weight decay of 3 10-4 and a batch size of 16. The initial learning rate was set to 2 10-5 and then minimized each epoch by the optimizer. A data augmentation technique was applied. Random patches from the input images were cropped and augmented horizontally and vertically. The size of the patches was fixed to 224×224 pixels. The feature extraction models were initialized with ImageNet pre-trained weights. Based on common practice in existing works, the dataset was divided based on the protocol 80% for training and 20% for testing randomly. For synthetically distorted datasets, the data were divided, referring to the original images to avoid overlapping. For all experiments, the 10-fold cross-validation was applied and the mean values of SROCC and PLCC were reported.

C. Evaluation and Discussion

Table II displays the results obtained in comparison to other existing works. The proposed model proved its superiority, especially on the LIVEFB and the KADID datasets, which are the largest datasets for synthetically and authentically distorted images, respectively. For small datasets, the proposed model presents similar or lower performances compared to existing works. Overall, the suggested model achieved competitive results. Considering the weighted average performance across all datasets, the recommended model achieved state-of-the-art performance for NR-IQA.

TABLE II. ACHIEVED PLCC AND SROCC IN COMPARISON TO EXISTING MODELS

Model	**LIVE**		**CSIQ**		**TID2013**		**KADID**		**CLIVE**		**KonIQ**		**LIVEFB**	
	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
P2P-BM [20]	0.958	0.959	0.902	0.899	0.856	0.862	0.849	0.840	0.842	0.885	0.872	0.598	0.526	
NSSADNN [21]	0.984	0.986	0.927	0.893	0.910	0.844	-	-	0.813	0.745	-	-	-	-
MetaIQA [22]	0.959	0.960	0.908	0.899	0.868	0.856	0.775	0.762	0.802	0.835	0.856	0.887	0.507	0540
TReS [23]	0.968	0.969	0.942	0.922	0.883	0.863	0.858	0.859	0.877	0.846	0.928	0.915	0.625	0.554
Proposed	0.972	0.983	0.945	0.918	0.886	0.867	0.856	0.861	0.873	0.877	0.934	0.921	0.628	0.557

A cross-dataset evaluation was conducted to further prove the performance of the model introduced. The training was performed on a dataset and the testing on another dataset without adaptation and finetuning. For the synthetic dataset, four common distortion types, including blur, JPEG2K, WN, and JPEG, were selected. Table III presents the achieved SROCC for the cross-dataset evaluation. The obtained results reveal that the proposed method outperformed the existing ones on five datasets while performing slightly lower on two other datasets. The proposed method proved its generalization power compared to other methods through this cross-dataset evaluation.

TABLE III. SROCC FOR CROSS-DATASET EVALUATION

Train dataset	**LIVEFB**	**CLIVE**	**KonIQ**	**LIVE**	**CSIQ**	**KADID**	**TID2013**
Test dataset	CLIVE	KonIQ	KADID	CSIQ	TID2013	LIVE	CLIVE
P2P-BM [20]	0.755	0.740	0.770	0.712	0.488	0.682	0.854
NSSADNN [21]	0.724	0.699	0.682	0.704	0.462	0.603	0.812
MetaIQA [22]	0.735	0.772	0.785	0.744	0.551	0.642	0.847
TReS [23]	0.713	0.733	0.786	0.761	0.562	0.681	0.835
Proposed	0.758	0.769	0.791	0.764	0.570	0.678	0.862

The proposed model demonstrates many advantages that ensure its efficiency for IQA. The hardware implementation of the lightweight CNN model on FPGA offers significant gains in terms of computational efficiency. FPGAs can run operations in parallel, resulting in faster processing times compared to traditional CPU-based implementations. Using FPGA-based hardware acceleration, the proposed approach enables real-time processing of IQA tasks. This is particularly advantageous in applications that require timely decisions or responses, such as surveillance systems or autonomous vehicles. FPGA-based implementations typically consume less power compared to CPU- or GPU-based systems, making them well-suited for mobile devices with limited battery capacity. This energy efficiency is crucial to extend the battery life of devices and reduce overall energy consumption. FPGAs offer a compact form factor that allows integration into mobile devices without significantly increasing their size or weight. This permits on-device image quality assessment capabilities without relying on cloud-based processing, thereby enhancing privacy and reducing latency. The proposed lightweight CNN model is designed to learn human visual behavior without prior knowledge of specific target visual behaviors. This self-learning ability allows the model to adapt and generalize well across diverse image quality assessment tasks and datasets.

However, the suggested method has a few disadvantages. Designing and implementing hardware-accelerated solutions on FPGAs can be complex and requires specialized knowledge in FPGA programming and hardware design. This complexity may pose challenges for developers without prior experience in FPGA development. Although FPGAs offer advantages in terms of parallel processing and energy efficiency, they have limited resources, such as logic cells, memory, and DSP blocks. Optimizing the lightweight CNN model to fit within these resource constraints without sacrificing performance can be a challenging task. Once programmed, FPGAs are not as flexible as software-based solutions and may require reprogramming or hardware modifications to accommodate changes or updates to the IQA algorithm. Ensuring the correctness and reliability of FPGA-based implementations

requires thorough verification and testing processes, which can be time-consuming and labor-intensive.

D. Ablation Study

An ablation study was carried out to investigate the impact of the proposed model. The proposed extraction backbones used for edge, texture, and semantic features extraction were explored to measure the influence of each of them. Table IV depicts the results achieved for each extraction backbone on the authentic datasets. Based on the results, it can be concluded that the performance of the suggested model, which incorporates multiple feature extraction backbones, exceeds that of a single feature extraction approach. The proposed method verified the idea of IQA by learning hierarchical features using different feature extraction backbones.

TABLE IV. ABLATION STUDY ON THE FEATURE EXTRACTION BACKBONES

	CLIVE		**KonIQ**		**LIVEFB**	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
Texture	0.831	0.842	0.857	0.868	0.545	0.521
Edge	0.859	0.851	0.821	0.828	0.601	0.503
Semantic	0.835	0.846	0.883	0.893	0.612	0.532
Proposed	0.873	0.877	0.934	0.921	0.628	0.557

Another ablation study was conducted to measure the impact of the proposed attention module on performance. For this, a performance comparison against simple fusion techniques was performed. Table V displays the results of the proposed fusion based on the attention module compared to other fusion techniques. The fusion technique based on the attention mechanism outperforms normal fusion techniques, such as addition and concatenation. This finding demonstrates the influence of the suggested contribution on the overall performance.

TABLE V. ABLATION STUDY OF THE FUSION TECHNIQUE

	CLIVE		**KonIQ**		**LIVEFB**	
	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC
Concatenation	0.869	0.843	0.901	0.876	0.621	0.533
Addition	0.839	0.827	0.893	0.891	0.610	0.515
Proposed	0.873	0.877	0.934	0.921	0.628	0.557

In summary, the ablation study demonstrates the influence of the proposed contributions on performance. The multi-backbone structure improved the feature extraction process, and the fusion mechanism collected the required features responsible for assessing image quality.

V. CONCLUSION

The hierarchy of visual quality degradation shifts with increasing distortion. It is challenging to find perfect images in reality when evaluating the quality of multi-distortion images. However, using deep learning, NR-IQA can not only identify additional forms of distortion, but can also be used in a wider range of contexts. The suggested method can detect deeper distortion types by separating interpretable hierarchical characteristics, including global semantics, edge, and texture. For this purpose, three backbones were deployed. To mitigate the impact of intricate correlations among hierarchical features on prediction results, the global and local contexts of intermediate features are computed using an attention-based feature fusion method. By reducing the disparity in fusion feature weights, it is possible to improve the understanding of the range of picture quality fluctuation. The experimental findings on seven major datasets disclose how accurate the model is at making predictions. The amount of data in the dataset has a significant impact on the accuracy of deep learning-based algorithms. Model training will not be sufficient if the dataset is small, and, as a result, the prediction outcomes will suffer. FPGAs tend to be more expensive than traditional CPUs or GPUs, both in terms of hardware cost and development time. This may limit the accessibility of the proposed approach to developers or organizations with budget constraints. Future research will concentrate on how to effectively extract characteristics from distorted images based on small datasets.

ACKNOWLEDGMENT

The authors gratefully acknowledge the approval and support of this research study by grant no. ENGA-2023-12-2071 from the Deanship of Scientific Research at Northern Border University, Arar, KSA.

REFERENCES

[1] G. Zhai and X. Min, "Perceptual image quality assessment: a survey," *Science China Information Sciences*, vol. 63, no. 11, Apr. 2020, Art. no. 211301, https://doi.org/10.1007/s11432-019-2757-1.

[2] M. Tan and Q. Le, "EfficientNetV2: Smaller Models and Faster Training," in *Proceedings of the 38th International Conference on Machine Learning*, Jul. 2021, pp. 10096–10106.

[3] Q. Huynh-Thu and M. Ghanbari, "Scope of validity of PSNR in image/video quality assessment," *Electronics Letters*, vol. 44, no. 13, pp. 800–801, Jun. 2008.

[4] S. Rani, Y. Chabrra, and K. Malik, "An Improved Denoising Algorithm for Removing Noise in Color Images," *Engineering, Technology & Applied Science Research*, vol. 12, no. 3, pp. 8738–8744, Jun. 2022, https://doi.org/10.48084/etasr.4952.

[5] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE Transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, Apr. 2004, https://doi.org/10.1109/TIP.2003.819861.

[6] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, Nov. 2003, vol. 2, pp. 1398-1402 Vol.2, https://doi.org/10.1109/ACSSC.2003.1292216.

[7] Z. Wang and Q. Li, "Information Content Weighting for Perceptual Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 5, pp. 1185–1198, Nov. 2010, https://doi.org/10.1109/TIP.2010.2092435.

[8] C. Galkandage, J. Calic, S. Dogan, and J.-Y. Guillemaut, "Full-Reference Stereoscopic Video Quality Assessment Using a Motion Sensitive HVS Model," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 31, no. 2, pp. 452–466, Mar. 2021, https://doi.org/10.1109/TCSVT.2020.2981248.

[9] H. T. R. Kurmasha, A. F. H. Alharan, C. S. Der, and N. H. Azami, "Enhancement of Edge-based Image Quality Measures Using Entropy for Histogram Equalization-based Contrast Enhancement Techniques," *Engineering, Technology & Applied Science Research*, vol. 7, no. 6, pp. 2277–2281, Dec. 2017, https://doi.org/10.48084/etasr.1625.

[10] L. Zhang, Y. Shen, and H. Li, "VSI: A Visual Saliency-Induced Index for Perceptual Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 23, no. 10, pp. 4270–4281, Jul. 2014, https://doi.org/10.1109/TIP.2014.2346028.

[11] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A Feature Similarity Index for Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, Dec. 2011, https://doi.org/10.1109/TIP.2011.2109730.

[12] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, Dec. 2014, https://doi.org/10.1109/TIP.2013.2293423.

[13] A. Maalouf, M.-C. Larabi, and C. Fernandez-Maloigne, "A grouplet-based reduced reference image quality assessment," in *2009 International Workshop on Quality of Multimedia Experience*, Jul. 2009, pp. 59–63, https://doi.org/10.1109/QOMEX.2009.5246975.

[14] I. P. Gunawan and M. Ghanbari, "Reduced-reference picture quality estimation by using local harmonic amplitude information," in *London Communications Symposium*, 2003, pp. 353–358.

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

[16] N. Ponomarenko *et al.*, "Image database TID2013: Peculiarities, results and perspectives," *Signal Processing: Image Communication*, vol. 30, pp. 57–77, Jan. 2015, https://doi.org/10.1016/j.image.2014.10.009.

[17] R. Vadhi, V. S. Kilari, and S. S. Kumar, "An Image Fusion Technique Based on Hadamard Transform and HVS," *Engineering, Technology & Applied Science Research*, vol. 6, no. 4, pp. 1075–1079, Aug. 2016, https://doi.org/10.48084/etasr.707.

[18] D. Ghadiyaram and A. C. Bovik, "Massive Online Crowdsourced Study of Subjective and Objective Picture Quality," *IEEE Transactions on Image Processing*, vol. 25, no. 1, pp. 372–387, Nov. 2015, https://doi.org/10.1109/TIP.2015.2500021.

[19] V. Hosu, H. Lin, T. Sziranyi, and D. Saupe, "KonIQ-10k: An Ecologically Valid Database for Deep Learning of Blind Image Quality Assessment," *IEEE Transactions on Image Processing*, vol. 29, pp. 4041–4056, 2020, https://doi.org/10.1109/TIP.2020.2967829.

[20] Z. Ying, H. Niu, P. Gupta, D. Mahajan, D. Ghadiyaram, and A. Bovik, "From Patches to Pictures (PaQ-2-PiQ): Mapping the Perceptual Space of Picture Quality," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 3572–3582, https://doi.org/10.1109/CVPR42600.2020.00363.

[21] B. Yan, B. Bare, and W. Tan, "Naturalness-Aware Deep No-Reference Image Quality Assessment," *IEEE Transactions on Multimedia*, vol. 21, no. 10, pp. 2603–2615, Mar. 2019, https://doi.org/10.1109/TMM.2019.2904879.

[22] H. Zhu, L. Li, J. Wu, W. Dong, and G. Shi, "MetaIQA: Deep Meta-Learning for No-Reference Image Quality Assessment," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, Jun. 2020, pp. 14131–14140, https://doi.org/10.1109/CVPR42600.2020.01415.

[23] S. A. Golestaneh, S. Dadsetan, and K. M. Kitani, "No-Reference Image Quality Assessment via Transformers, Relative Ranking, and Self-Consistency," in *2022 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, Waikoloa, HI, USA, Jan. 2022, pp. 3989–3999, https://doi.org/10.1109/WACV51458.2022.00404.