IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 25 May 2024, accepted 18 June 2024, date of publication 24 June 2024, date of current version 1 July 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3418375

RESEARCH ARTICLE

TQP: An Efficient Video Quality Assessment Framework for Adaptive Bitrate Video Streaming

MUHAMMAD AZEEM ASLAM[1,2], XU WEI[2], NISAR AHMED[3], GULSHAN SALEEM[4], ZHU SHUANGTONG[2], YIMEI XU[1], AND HU HONGFEI[1]

[1]School of Information Engineering, Xi'an Eurasia University, Xi'an, Shaanxi 710065, China
[2]Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, Jilin 130033, China
[3]Department of Computer Engineering, University of Engineering and Technology at Lahore, Lahore, Punjab 54890, Pakistan
[4]Department of Computer Science, Lahore Garrison University, Lahore, Punjab 54000, Pakistan

Corresponding author: Muhammad Azeem Aslam (azeem@eurasia.edu)

This work was supported in part by Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, China; in part by the School of Information Engineering, Xi'an Eurasia University, Xi'an, Shaanxi, China; and in part by Chinese Academy of Sciences President's International Fellowship Initiative under Grant 2024PVB0036.

ABSTRACT The increasing popularity of video streaming services and the widespread accessibility of high-speed internet underscore the importance of delivering cost-effective and seamless streaming experiences. Shared internet connections may lead to varying speeds, impacting Quality of Experience (QoE). Rate adaptation techniques aim to ensure smooth video transmission, but overly optimistic adaptations can compromise user experience. Objective video quality assessment is crucial for efficient rate adaptation to ensure smooth QoE. This research proposes a novel method incorporating temporal channel shifting into Convolutional Neural Networks (CNN) for video quality assessment while maintaining the computational simplicity of a 2D CNN model. The proposed approach relies on the EfficientNet architecture, initially pre-trained on quality-aware images, and fine-tune it using datasets of rate-adaptive videos. The model is trained and evaluated on two benchmark datasets, namely "Waterloo sQoE III" and "LIVE Netflix II," which consist of rate-adaptive videos annotated with subjective quality scores. Experimental results encompass the evaluation of Pearson, Spearman, and Kendall correlation coefficients, along with the computation time ratio for the proposed approach. The outcomes reveal competitive scores of 0.795, 0.652, 0.772, and 0.216 for the "Live Netflix II dataset" and 0.782, 0.713, 0.721, and 0.230 for the "Waterloo sQoE III dataset." Our proposed method, compared to 24 approaches for "Waterloo sQoE III" and 25 for "LIVE Netflix II," attains the highest correlation scores while maintaining near-real-time processing efficiency. These results affirm the efficacy of our approach in accurately predicting human judgment (QoE) with computational efficiency.

INDEX TERMS Video quality, image quality assessment, rate adaption, video streaming, quality of experience, QoE.

I. INTRODUCTION

The popularity of video applications has increased due to the rapid advancements in digital multimedia devices and the widespread availability of inexpensive, high-speed internet. This has resulted in a notable rise in internet traffic that is attributed to multimedia data, specifically images and videos [1], [2], [3]. Smartphones, equipped with high-definition cameras and versatile connectivity options such as Wi-Fi and cellular networks, exemplify these multifunctional devices contributing to the escalating demand for multimedia content. Cisco's annual internet study reveals that video data constitutes a staggering 82% of internet traffic [4]. This trend intensifies the competition among video distribution systems and streaming services [5], where the success of these platforms is intricately linked to the end

The associate editor coordinating the review of this manuscript and approving it for publication was Miaohui Wang.

© 2024 The Authors. This work is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 License.
For more information, see https://creativecommons.org/licenses/by-nc-nd/4.0/

user's Quality of Experience (QoE) [6]. Consequently, there is a critical need for robust quality assessments to gauge users' QoE and optimize the distribution of video content effectively [7], [8].

The shared nature of internet connections introduces variability in connection speeds, heavily influenced by network traffic generated by other users. In the context of adaptive bitrate streaming, where seamless content delivery necessitates adaptive bitrate adjustments [9], ensuring perceptual quality becomes pivotal for overall QoE. Overly optimistic rate adaptations may lead to a reduction in throughput requirements, adversely affecting QoE [9], [10]. To address this challenge, perceptual quality assessment of video streams becomes essential, guiding efficient rate adaptation by optimizing network channels based on improved perceptual quality assessment models.

Objective quality assessment methods fall into three categories: Full-Reference (FR), Reduced-Reference (RR), and No-Reference (NR) [11]. FR methods compare original and reproduced content for accurate assessment, while RR methods use additional features transmitted with the original content [12]. NR methods, though challenging, operate blindly on the original content, making them more practical in various applications where reference content is unavailable [1]. Developing accurate and efficient NR Video Quality Assessment (VQA) methods is crucial for widespread applicability.

Despite advances in image quality assessment, videos have received less attention in the NR quality assessment domain, primarily due to the temporal complexity introduced by video's dynamic nature [13], [14]. For streaming videos incorporating rate adaptation, considering the temporal element in quality evaluation becomes crucial, capturing scene transitions and temporal artifacts. Recent efforts in developing objective algorithms for predicting visual quality have involved traditional approaches leveraging natural scene statistics or other features for regression algorithm training [15].

Perceptual quality evaluation is conducted using both subjective and objective methods [7], [11]. While subjective assessment involves human observers assigning quality scores [16], objective methods offer a cost-effective alternative across multimedia applications [17], [18], [19], [20], [21]. Objective methods are classified as Full-Reference (FR), Reduced-Reference (RR), or No-Reference (NR) [11]. Despite their limitations, NR methods are useful in situations where reference content is not available [1]. Video quality assessment is challenging due to its dynamic nature [13], [14], and recent efforts aim to incorporate temporal elements for accurate evaluation [15]. The rising trend in deep learning, particularly Convolutional Neural Networks (CNNs), has gained prominence in VQA research [22]. Two approaches exist in VQA: one incorporating temporal elements and another that does not. While CNN-based approaches provide frame-by-frame predictions, three-dimensional CNNs offer spatiotemporal learning by considering multiple frames as input. However, these approaches often incur high computational costs, limiting their deployment on end devices and real-time online video understanding [23], [24]. This research addresses VQA for rate-adaptive video streaming, leveraging the EfficientNet 2D CNN architecture pre-trained on quality-aware datasets BIQ2021 and KonIQ-10K. Introducing channel shift for temporal feature propagation, the proposed approach balances efficiency, accuracy, and generalization capabilities. The channel shift is implemented in the residual connection of the architecture, and scalability within the EfficientNet family allows for enhanced predictive performance. Additionally, a quality-aware loss function, combining mean squared error, mean absolute error, and Spearman's rank order correlation coefficient, contributes to a comprehensive evaluation of video quality.

This study makes significant contributions, including proposing an efficient VQA model tailored for adaptive bitrate video streaming, quality-aware pre-training using BIQ2021 and KonIQ-10K datasets, and introducing a novel quality-aware loss function. The subsequent sections delve into the framework, experimental setup, results, and implications, providing a thorough exploration of the proposed framework's capabilities within adaptive bitrate video streaming scenarios.

II. RELATED WORK

VQA for video streaming applications has been approached through various methods, categorized into network/client-side statistics and perceptual content-based approaches. QoS-oriented approaches, such as those based on bitrate, latency, and rebuffering time [25], [26], [27], [28], [29], offer fast but less accurate solutions, suitable for simplistic quality predictions and limited computational complexity devices. Hybrid approaches [30], [31], [32], [33] integrate network statistics and spatiotemporal characteristics, enhancing prediction performance at the expense of computational complexity. On the other hand, content-based approaches prioritize perceptual aspects, showcasing high correlation with human judgment but often demanding increased computational resources.

Two popular categories of content-based approaches are Image Quality Assessment (IQA)-based and VQA-based. IQA-based methods like BRISQUE [34], NIQE [35], and DeepEns [1] predict video quality by treating frames as separate images. Nevertheless, the determination of the optimal frames per second and the oversight of temporal characteristics with 2D CNNs pose challenges, restricting their applicability for adaptive bitrate videos [1], [34], [35].

Content-based video quality assessment methods, like VIIDEO [36] and V-BLIINDS [37], introduce no-reference algorithms to predict video quality based on statistical and perceptual features. While effective, these methods tend to be computationally expensive, limiting real-time deployment in streaming applications.

Several approaches, including UGC-VQA [38], PVQ [39], VSFA [40], and FAST-VQA [41], address content variation and attempt to overcome computational inefficiencies.

However, challenges persist in terms of computational complexity, model adaptability, and the representation of perceptual video quality.

Research efforts like StarVQA [42] propose attention-based approaches for video quality assessment, leveraging transformer networks. While providing good predictive performance, computational efficiency remains slightly below real-time standards.

Despite these advancements, a research gap exists in achieving a balance between predictive performance and computational efficiency in video quality assessment. Existing approaches either prioritize one aspect at the expense of the other or encounter challenges in adapting to real-time processing requirements. This research aims to bridge this gap by proposing an efficient video quality assessment framework tailored for adaptive bitrate video streaming, ensuring both accuracy and computational efficiency.

III. PROPOSED METHODOLOGY

Three-dimensional CNNs are designed to circumvent this limitation by processing multiple video frames at once and capturing the temporal relationship between the adjacent frames [43], [44], [45], [46]. This capability of the 3D CNNs (such as C3D [47] or I3D [48]) makes them unusually complex by increasing the required number of training parameters to the number of frames processed at a time [49]. This study, on the other hand, proposes a Temporal Quality Predictor (TQP) framework, which combines various concepts from existing studies in order to carry out temporal modeling without introducing extra levels of complexity [1], [50], [51]. Figure 1 depicts the whole process that the TQP employs in order to make its quality predictions for videos.

A. MODELING APPROACH

Our modeling approach is based on the assumptions that similar to spatial redundancy, the frames in a video stream are also highly redundant and the change in content in subsequent frames is relatively low. This assumption has led us to explore channel shifting in the temporal dimension in which only a portion of the information is shared with adjacent frames to learn the temporal characteristics. The proposed approach ensures an increase in model capacity and minimization of data shifting which leads to efficient and accurate models. This strategy ensures that spatial learning is not heavily impaired and the incorporation of temporal learning is enough to model temporal feature learning.

The model employs a strategy of using off-the-shelf, pre-trained CNN to perform temporal feature extraction in residual connections using temporal channel shifting. EfficientNet [37] is used as a pre-trained model because of its small size, good generalization and scalability, but any CNN model with residual connections can be modified to perform channel shifting. In addition, the framework includes a pre-training strategy, temporal feature extraction, and quality estimation via regression using spatiotemporal features extracted by the Temporal Quality Predictor (TQP). Each of these stages is described in greater detail in the following subsections.

The channel shifting performs information sharing between subsequent frames and is demonstrated using 2D convolution operation which is given by Eq. 1 & 2.

$$y[a,b] = \Sigma_{m=-\infty}^{\infty}\Sigma_{n=-\infty}^{\infty} h[m,n]\cdot x[a-m,b-n] \quad (1)$$

$$\begin{aligned} y[m,n] &= x[m,n] * h[m,n] \\ &= \Sigma_{m=-\infty}^{\infty}\Sigma_{n=-\infty}^{\infty} x[I,j]\cdot h[m-i,n-j] \end{aligned} \quad (2)$$

Here x and y are the input and output image, h is the kernel matrix with m and n indices and (i, j) provide the pixel location. The convolution operation is performed for all possible values and the padding is decided as per the user's specifications. In case of a 1D input f and kernel h the convolution $Conv(f, g)$ requires a weight vector g with size three (g_1, g_2, g_3) to be convolved with input f. The convolution for this scenario can be given by Eq. 3.

$$fi = g_1 f^{-1} + g_2 f^{0} + g_3 f^{+1} \quad (3)$$

In order to perform convolution of the video frame through shifting and multiplication, the input channels can be shifted by -1 and $+1$. The shift operation performs information transfer with its neighboring frames ($f_{i-1}, f_i and f_{i+1}$) and the operation can be performed as described in Eq. 3.

$$\begin{bmatrix} A_i^{-1} = A_{i-1} \\ A_i^{0} = A_i \\ A_i^{+1} = A_{i+1} \end{bmatrix}$$

Figure 2 demonstrate the channel shifting in temporal dimension by using eight channels and three consecutive frames. The channels learned from different frames are represented by different colors to highlight them during channel shift taking place in the temporal dimension. Figure 2(a) illustrate the original channel sequence in consecutive frames whereas Figure 2(b) illustrates the channel sequence after the application of shift. Channel 1 in Figure 2(b) demonstrates left temporal shift which indicates shifting by -1 shift in frames whereas channel 2 in Figure 2(b) demonstrates right temporal shift by $+1$ shift in frames taking place in channel 2. The rest of the channels in Figure 2(b) demonstrate unshifted frames.

In our study, we have adopted a channel shift of 1/8 channels, a strategy aligned with established methodologies such as TSM. This approach offers several advantages, including reduced computational overhead by shifting only a fraction of channels, as well as the retention of spatial learning ability, as highlighted by Lin et al. [51]. Furthermore, this choice aligns with the understanding that temporal frames typically exhibit less variation in information compared to spatial blocks. We believe this approach strikes a balance between computational efficiency and preserving important spatial-temporal features essential for our model's performance.

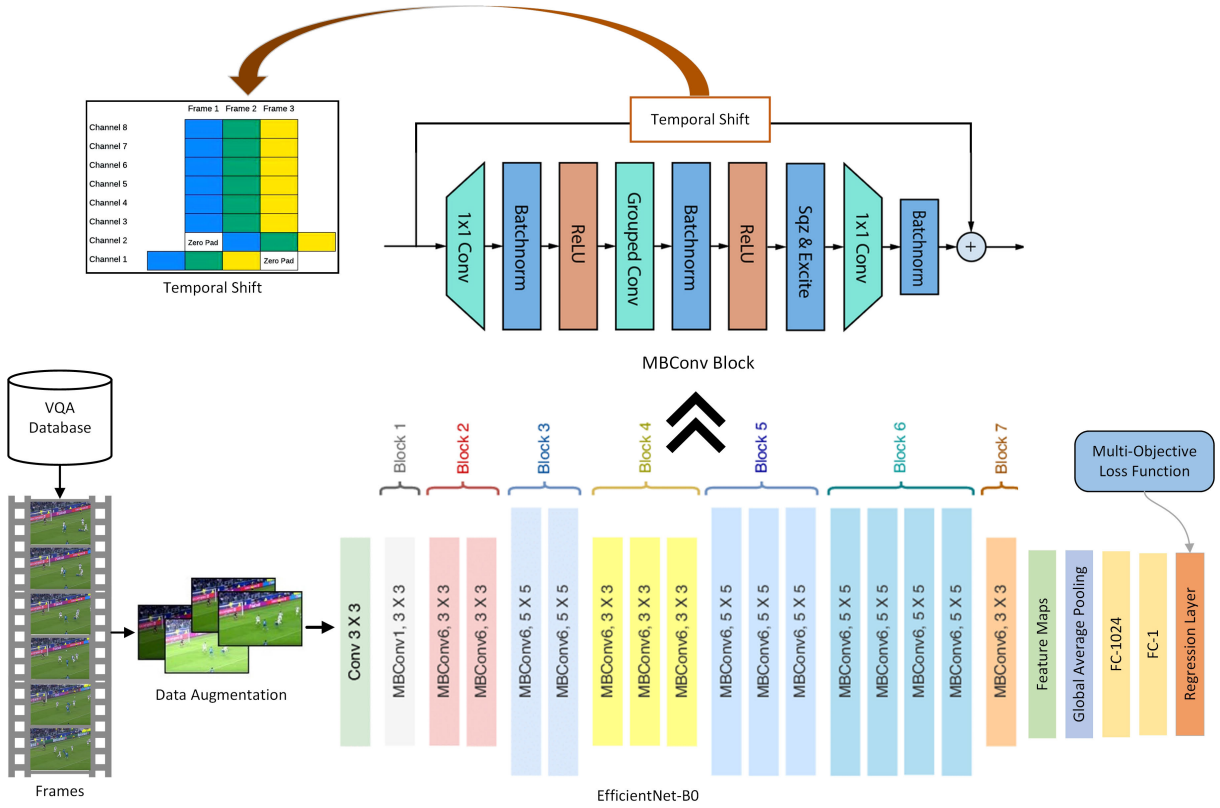


FIGURE 1. **Overall framework of Temporal Quality Predictor (TQP).**

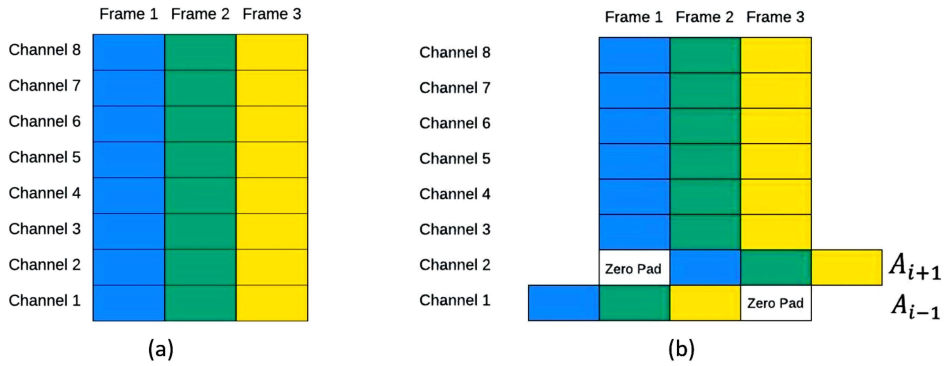


FIGURE 2. **Graphical depiction of channel shifting in temporal dimension.**

B. QUALITY PREDICTOR

In order to perform quality prediction, spatiotemporal feature extraction is performed using EfficientNet architecture with channel shifting. The video stream is converted into individual frames which are saved on disk and provided sequentially to the model for spatiotemporal feature learning. The frames are augmentation as part of pre-processing before providing them to the spatiotemporal feature extractor to increase effective dataset size and allow the model to learn variations resulting in improved generalization. The information extracted in terms of learned features is transferred from each frame to its subsequent frames after pre-processing and data augmentation. Fig. 2 depicts the channel shift in both directions of the temporal dimension and therefore allows the model to learn feature maps via the EfficientNet model and the use of temporal channel shift.

1) EFFICIENTNET ARCHITECTURE

EfficientNet is a family of CNN models introduced by [47] and is among the most efficient models (i.e. requiring the least FLOPS for inference) that reach State-of-the-Art accuracy on

both ImageNet and common image classification & transfer learning tasks. The baseline architecture termed EfficientNet-B0 is similar to MnasNet, which reached near-SOTA with a significantly smaller model. Model scaling is performed through a heuristic way to perform compound scaling in depth, width, and resolution dimensions. The scaled-up versions are EfficientNet-B1 to B7 which represents a good combination of efficiency and accuracy on a variety of scales. Such a scaling heuristics (i.e. compound-scaling) allows the efficiency-oriented base model (B0) to surpass models at every scale while avoiding extensive grid-search of hyperparameters.

The architecture also includes various design elements that improve efficiency. Squeeze-and-Excitation (SE) blocks allow the network to dynamically adjust the channel-wise feature responses, resulting in improved accuracy. Stem blocks are used at the beginning of the network to reduce the spatial resolution of the input image and make the network more efficient. The swish activation function is used instead of traditional activation functions such as ReLU to improve the model's accuracy and efficiency. The SE block is inserted into the network between the convolutional layers and the activation function. The use of SE blocks has been shown to significantly improve the accuracy and efficiency of the network.

The residual connection is modified to implement a channel shift, which leads to the extraction of both temporal and spatial features when using EfficientNet with a temporal channel shift. To mitigate the latency overhead of the model and overcome the compute cost of the framework, the model propagate 1/8 of the features to the adjacent frames. The information in the current frame is shifted along the temporal dimensions and the learned parameters are shared with connected or neighboring frames in the channel shift.

In order to perform quality prediction, the last fully connected layer has single neuron followed by regression layer. The loss function used for model training is a quality-aware loss function which is described in the III-C. The model performs spatial feature learning whereas temporal feature learning is performed through channel shifting implemented in the residual block. As the temporal channel shift introduces channels in the next frame for each frame and therefore can perform long-range temporal modeling through cascading effect due to the transfer of information from each frame to its neighboring frames.

C. LOSS FUNCTION

MSE is the most frequently used loss function for training quality assessment models. Ahmed et al. [1] demonstrated empirically that MSE is the best loss function for training image quality assessment models. In contrast, Hosu et al. [52] claimed that MAE is the most appropriate loss function for assessing image quality. However, we are attracted to training our model using a quality-aware loss function rather than just MSE and MAE.

In order to achieve this goal, a loss function is formulated by integrating multiple error metrics, including MSE and MAE, along with a differentiable approximation of Spearman's Rank Order Correlation Coefficient (SROCC). The loss function can be obtained using Eq. 4.

$$\text{Loss} = \lambda_1 \cdot \text{MSE} + \lambda_2 \cdot \text{MAE} + \lambda_3 \cdot (1 - \text{SROCC}) \quad (4)$$

MSE quantifies the average squared difference between predicted and ground truth quality scores, defined by Eq. 5.

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 \quad (5)$$

where Y_i and $\hat{Y}_i$ denote the ground truth and predicted scores. Similarly, MAE measures the average absolute difference, given by Eq. 6.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|Y_i - \hat{Y}_i| \quad (6)$$

To address the non-differentiability of traditional SROCC calculation, a differentiable approximation is introduced. In this approximation, the softmax function is employed to convert predicted and ground truth scores into continuous and differentiable ranks. The softmax ranks R_i are calculated by Eq. 7 for both predicted and ground truth scores. These softmax ranks are then used to compute the differentiable approximation of SROCC. This formulation allows for efficient gradient-based optimization during CNN training while simultaneously capturing both magnitude and order-related errors in image quality predictions.

$$R_i = \frac{e^{Y_i}}{\sum_{j=1}^{n} e^{Y_j}} \quad (7)$$

In Eq. 4, the coefficients λ_1, λ_2, and λ_3 control the relative importance of each loss component. The differentiable SROCC component is $1 - \text{SROCC}$ to frame it as a loss term (i.e., minimizing it is desirable). The MSE and MAE components serve as traditional error metrics, capturing the difference between the predicted and ground truth scores in terms of squared and absolute differences, respectively.

The rationale behind the formulation of the multi-objective loss function of Eq. 4 to consider the magnitude and the direction of the error as well as the monotonic relation between the predicted and ground-truth values. On one hand, incorporating both MSE and MAE, the loss function aims to minimize the squared differences between the predicted and ground truth quality scores while also considering the absolute differences. This allows CNN to not only focus on minimizing the overall error but also pay attention to the accuracy of the predictions. SROCC component measures the monotonic relationship between the predicted and ground truth ranks. By including it in the loss function, the model is encouraged to learn to predict not only the quality scores but also their relative order or ranking.

To determine the values of hyperparameters λ_1, λ_2, and λ_3, a grid search was conducted using values of 1/2, 1, and 2.

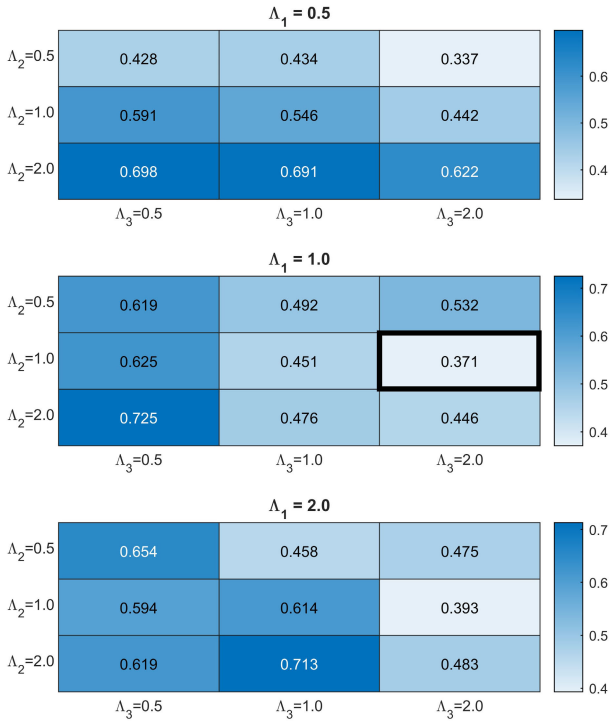


FIGURE 3. Loss values at various grid points.

This exploration involved 27 distinct configurations, each trained for 10 epochs. Figure 3 provides the values of loss at the end of training for various grid points indicating that lowest loss value of 0.371 is obtained at $\lambda_1 = 1$, $\lambda_2 = 1$ and $\lambda_3 = 2$. Subsequently, Equation 8 was derived as a simplified form of the loss function utilized during model training.

$$\text{Loss} = \text{MSE} + \text{MAE} + 2 \cdot (1 - \text{SROCC}) \tag{8}$$

D. PRE-TRAINING

Training of a CNN usually requires a large number of labeled images which is a challenging task in many cases such as VQA. Transfer learning is a popular method where a model developed for one task is reused as a starting point to model a different task. This method presents itself as an excellent opportunity and eliminates the need for a vast amount of computational time and training data. In order to take benefit of this approach and eliminate the need for a large amount of labeled training data, we have performed model pre-training on image quality assessment database BIQ2021 [16]/KonIQ-10K [52]. As the MOS of both dataset follows different scale and therefore the MOS is normalized before combining and using both datasets to perform model pre-training. The Eq. 9 is used to perform normalization of the MOS. This pre-training using min-max normalized MOS score allows the TQP to learn the quality aware priors and therefore improve the subsequent VQA performance.

$$\overline{MOS}_i = \frac{MOS_i - min(MOS)}{max(MOS) - min(MOS)} \tag{9}$$

where i is the index of an element in MOS and $\overline{MOS}$.

E. DATA AUGMENTATION

Data augmentation is a commonly used technique in training deep learning models and is required to serve various purposes. It increases the effective dataset size by introducing artificially generated training examples. This increased dataset size allows the model to reduce dependency on a larger dataset. It improves the model generalization by introducing diverse variations in the training data which sometimes doesn't occur in the data. It allows the training process to mitigate the overfitting as the model may become too specialized on the training instances and the introduction of perturbations allow it to learn more general features. These generalized features make the model more robust to real-world variations and therefore lead to improved performance.

To train the TQP, we have introduced horizontal flip which makes the model invariant to horizontal flipping. Vertical and horizontal translation and left and right rotation up to 5^0 are introduced for additional variations in the image. Moreover, as the model accepts an input size of 224×224 and the spatial resolution of the video is different from this size, therefore random cropping is performed to select a random patches of the frame equal to the input size. These variations allow the model to learn a generalized quality aspect of the image and focus less on the content in the image (frame).

F. TRAINING OPTIONS

In order to train the model, we used Adam optimizer and trained it with a batch size of 16 images. An initial learning rate of 3×10^{-3} is used with a piece-wise learning rate in which the learning rate is halved after 30, 20, and then 10 epochs. The model is trained for 300 epochs with a validation check to avoid overfitting. The training is stopped if the validation loss stops decreasing after three validation checks.

IV. EXPERIMENTS

In this section, we present the details of the experimental setup used to evaluate the proposed approach for VQA. The execution environment, description of the dataset, and evaluation metrics for the experiments are discussed followed by model evaluation and comparison with existing approaches.

A. EXECUTION ENVIRONMENT

The experiments were conducted using Matlab ® 2022b on a Windows 10 equipped Dell T3610 workstation. The workstation is powered by an Intel ® Xeon ® Processor E5-2687W v2 and has 32GB of RAM. To enhance performance and reduce latency, the training dataset, operating system and Matlab were installed/placed on a SATA SSD. The system was equipped with a GeForce RTX 3060 GPU with 12GB GDDR6 memory, providing ample computational power for the experiments. This setup ensured a robust and efficient

execution environment for the research, enabling smooth processing and analysis of the data.

B. DATASET DESCRIPTION

In order to perform VQA, a subjectively scored dataset is a prime requirement as training of a supervised learning model is performed using labelled examples. Several studies have targeted the QoE issue of streaming videos, but they lack aspects of real-world video streaming systems such as actual network measurements and client-driven adaptation strategies. Therefore, we have relied on the "LIVE-NFLX-II Subjective Video QoE Database" [53] and "Waterloo sQoE Database-III" [54] which are two comprehensive subjective QoE databases containing rate adaptive videos.

1) LIVE-NFLX-II SUBJECTIVE VIDEO QOE DATABASE

The LIVE-NFLX-II [53] dataset contains 420 videos that were evaluated by 65 subjects. The dataset provides 9,750 continuous-time and retrospective subjective opinion scores. The dataset provides instantaneous QoE-based subjective opinion scores in a continuous-time scenario and overall video quality score for a retrospective scenario. We are dealing with long-term quality assessment scenarios in this study and are therefore using retrospective subjective scores only. Figure 4 provides a depiction of nine sample videos from the LIVE-NFLX-II database. The distribution of Mean Opinion Score (MOS) for the dataset is provided in Figure 5. The dataset is created by generating video content from 15 videos streamed under seven different network conditions and four different adaptation strategies. It is a large-scale dataset in terms of video content, encoding, and adaptive streaming. The dataset can be used for VQA, perceptual video coding, incorporation of buffer and network conditions, and client-driven adaptation research. The seven network conditions are actual network traces from the HSDPA dataset which represents a challenging 3G mobile networks scenario. Moreover, four client adaptation strategies are used to generate content on the basis of rate, buffering, and quality, whereas the video content itself covers a range of content categories. The content characteristics span a large variety

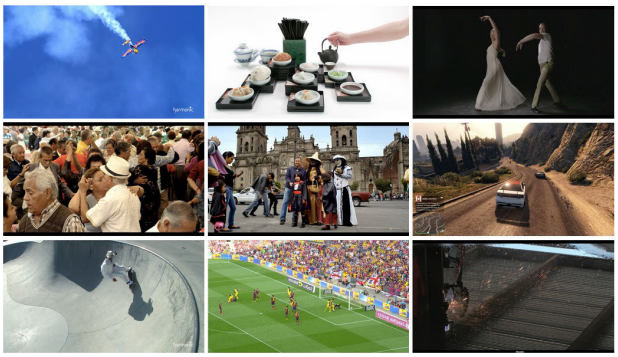


FIGURE 4. Sample video frames depicting the content diversity in the LIVE-NFLX-II database.

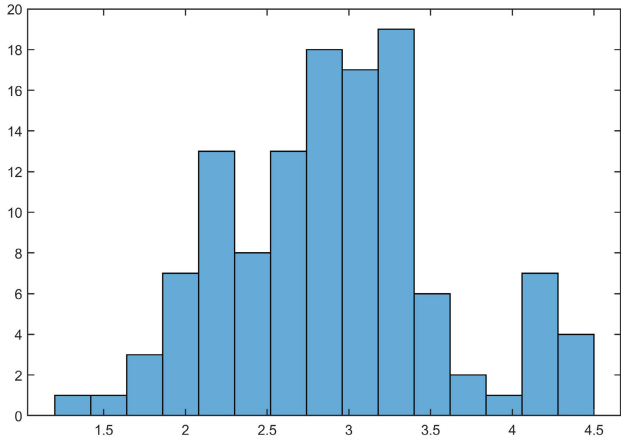


FIGURE 5. Distribution of MOS in the LIVE-NFLX-II database.

including natural and animation video content, fast/slow motion scenes, light/dark scenes, and low and high texture scenes.

To accurately model video streaming, the database utilizes actual network measurements and a pragmatic client buffer simulator, going beyond simplistic network and buffer occupancy models. The database captures various aspects of streaming adaptation, including video quality fluctuations, re-buffering events of different duration and frequencies, spatial resolution changes, and diverse bitrate/quality levels across different video content types. By incorporating multiple network traces and adaptation strategies, the database offers a comprehensive representation of real-world video streaming scenarios.

2) WATERLOO SQOE DATABASE-III

The Waterloo Streaming Quality-of-Experience Database-III (sQoE-III) [54] is a comprehensive dataset designed for VQA. It comprises of 20 RAW HD reference videos and 450 simulated streaming videos, with an average duration of 13 seconds. To ensure the generation of meaningful and representative test videos, a series of Dynamic Adaptive Streaming over HTTP (DASH) experiments were conducted. The relevant streaming activities were recorded, and the streaming sessions were reconstructed using video processing tools. The streaming sessions were generated using six adaptive streaming algorithms: rate-based, BBA, AIMD, Elastic, QDASH, and FESTIVE. These algorithms were evaluated under 13 diverse bandwidth conditions. A total of 34 subjects participated in the evaluation process, scoring the quality of each video sequence using a numerical quality scale ranging from 0 to 100. Figure 6 provides the depiction of nine sample video frames from the Waterloo dataset. The distribution of MOS for the dataset is provided in Figure 7. The uniqueness of the SQoE-III database lies in its realistic and diverse nature. Unlike existing databases that often contain hand-crafted test sequences, the SQoE-III database provides a larger and more representative collection of 450 streaming videos. These videos are created from diverse

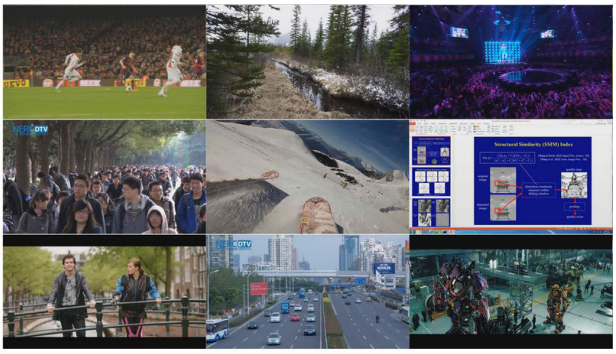


FIGURE 6. Sample video frames depicting the content diversity in the database.

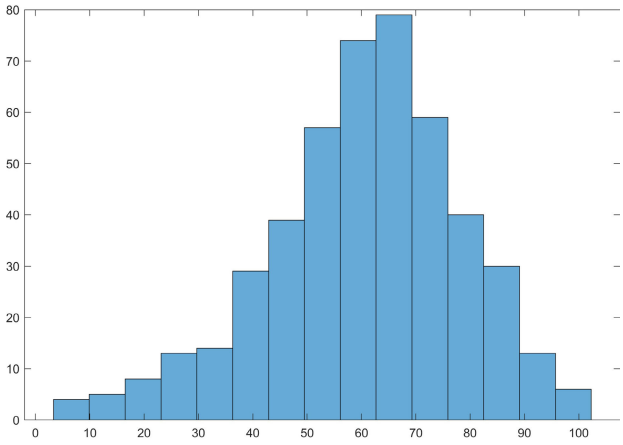


FIGURE 7. Distribution of MOS in the Waterloo database.

source content and encompass a wide range of distortion patterns. Additionally, the dataset includes recordings of streaming sessions using six adaptation algorithms with distinct characteristics under 13 representative network conditions. The quality of all streaming videos was assessed by the 34 subjects.

C. EVALUATION METRICS

1) PEARSON LINEAR CORRELATION COEFFICIENT(PLCC)

PLCC measures the linear correlation between the predicted quality scores and the ground truth scores. It ranges from −1 to 1, with 1 indicating a perfect positive correlation, −1 indicating a perfect negative correlation, and zero for no correlation. A higher PLCC value signifies better performance. The formula for calculating PLCC is provided Eq. 10.

$$\text{PLCC} = \frac{\sum_{i=1}^{n}((x_i - \bar{x})(y_i - \bar{y}))}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{10}$$

2) SPEARMAN RANK-ORDER CORRELATION COEFFICIENT (SROCC)

SROCC measures the monotonic relationship between the predicted quality scores and the ground truth scores, disregarding any linear correlation. Similar to PLCC, SROCC ranges from −1 to 1, with higher values indicating better performance. The formula for calculating SROCC is provided in Eq. 11.

$$\text{SROCC} = 1 - \frac{6\sum_{i=1}^{n}(d_i)^2}{n(n^2-1)} \tag{11}$$

In the formula of Eq. 11 & Eq. 10, n represents the number of samples, x_i and y_i represent the image quality scores for the i_{th} sample, and $\bar{x}$ and $\bar{y}$ represent the means of the quality scores. d_i represents the difference in ranks between corresponding pairs of quality scores.

3) KENDALL RANK-ORDER CORRELATION COEFFICIENT (KRCC)

KROCC also measures the monotonic relationship between the predicted quality scores and the ground truth scores, but it emphasizes concordant and discordant pairs. A higher KROCC signifies better performance. The formula for calculating KROCC is provided in Eq. 12.

$$KROCC = \frac{C - D}{\sqrt{(C + D + T) \cdot (C + D)}} \tag{12}$$

In this formula, C represents the number of concordant pairs, D represents the number of discordant pairs, and T represents the number of ties in the dataset. The formula calculates the KROCC (τ) by dividing the difference between the number of concordant pairs and the number of discordant pairs by the square root of the product of $(C + D + T)$ and $(C + D)$.

4) COMPUTATION TIME RATIO (CTR)

The CTR is a metric that represents the ratio of the computation time to the duration of the video being processed. It indicates the efficiency of the model in performing the task relative to the duration of the video. The formula to calculate the computation time ratio can be expressed as Eq. 13:

$$CTR = ComputationTime/VideoDuration \tag{13}$$

where,

Computation Time: refers to the time taken by the system or model to perform the computation or processing of the task.

Video Duration: represents the total duration or length of the video being processed.

It's important to note that the computation time and video duration are measured consistently using the same unit of time (e.g., seconds) to ensure accurate comparison and calculation of the ratio. Additionally, it's essential to note that the computation time is calculated with a single CPU only without the support of parallel processing or GPU.

D. MODEL EVALUATION

To evaluate the performance of proposed framework (TQP), the dataset is partitioned into train/test split. The training set contains 80% of the video streams whereas the remaining 20% are reserved for model testing. This partitioning ensured that the model was trained on a sufficient amount of data

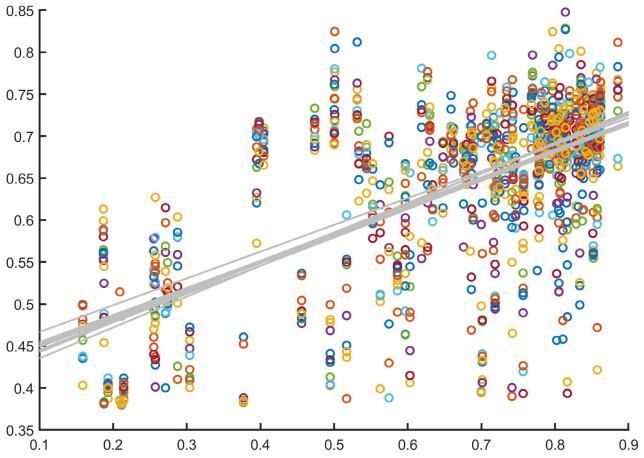


FIGURE 8. Scatter-plot of ground-truth vs predicted values for 10 iterations (the regression line for the model is plotted for each iterations).

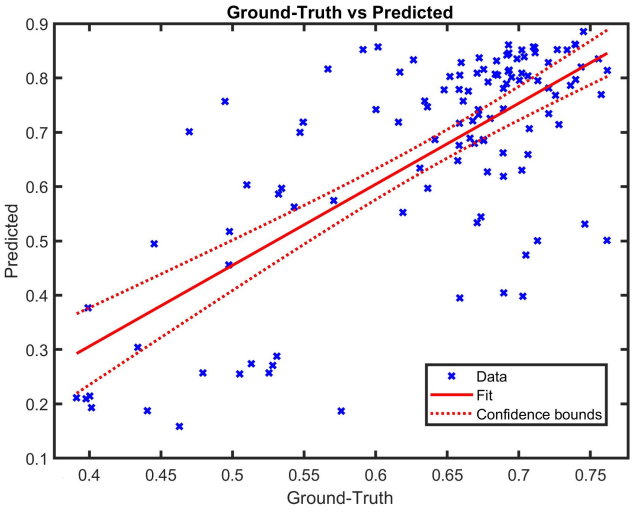


FIGURE 9. Scatter-plot of ground-truth vs average predictions along with the confidence interval.

while also allowing for an unbiased evaluation on unseen video streams.

During the testing phase, the quality predictions are performed for 10 iterations. In each iteration, the model randomly cropped a section of the input video sequence with a spatial resolution of 224 × 224. This random cropping approach helped capture diverse spatial information from different regions, providing a robust evaluation of the model's ability to generalize to various regions. Figure 8 provide the scatter plot and fitting line for 10 iterations for LIVE Netflix II dataset. For each iteration, the model produced a predicted quality score for the cropped image parch. To obtain a final predicted quality score for the entire video sequence, we averaged the quality predictions from the 10 patches (iterations). This averaging process accounted for the variability introduced by the random cropping and provided a more reliable estimate of the overall video quality. The scatter plot between ground-truth and the final predicted quality score is provided in Figure 9 with the regression line and its confidence interval. The residual plot is an important tool for evaluating the performance and reliability of a video quality prediction model. It is constructed by plotting the differences (residuals) between the predicted quality scores and the corresponding ground truth quality scores. Each data point on the plot represents a video sample, with the x-axis representing the predicted quality score and the y-axis representing the residual. It allows, assessment of model's predictions with the actual observed quality scores and provides insights into the model's accuracy and potential sources of error. It has the ability to reveal any systematic patterns or trends in the model's predictions. A well-fitted and accurate model should exhibit a random scattering of residuals around the zero line, indicating that the predictions are unbiased and capture the true underlying quality of the videos. On the other hand, systematic patterns or trends in the residuals may suggest the presence of model bias or limitations.

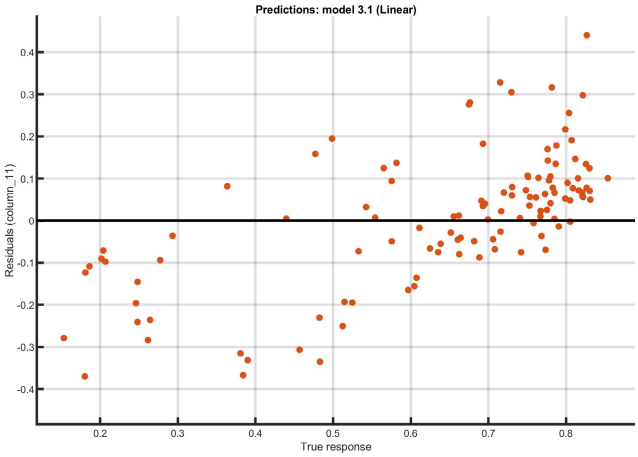


FIGURE 10. Residual plot with true predictions at the x-axis and residual value at the y-axis.

The residual plot of Figure 10 indicates that the data does not have a strong trend but the distribution of the points is negative for the first half of the predictions and positive for the latter half and therefore indicates an apparent weakness in the model. The comparison of the predictive performance with existing approaches indicated that the proposed model provides the highest correlation with the ground-truth and the bias in the model is the result of a smaller dataset size. As the number of samples in the LIVE Netflix II dataset is 420 and for Waterloo sQoE -III dataset are 450 which are not representative enough to learn a good generalization. The residual analysis was performed during the design of the model and several considerations were explored and the proposed TQP has provided the best performance and therefore warrants for an increase in training dataset size to improve the model's generalization. The response plot of the proposed approach is provided in Figure 11 which provide the scatter plot of ground-truth and predicted values for each

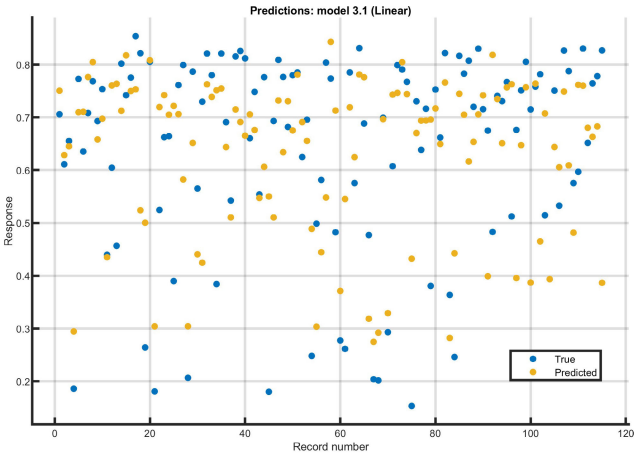


FIGURE 11. Response plot of ground-truth and predicted values for each sample of test set.

record in the test set. The plot is useful in visualizing the actual and predicted quality scores for each record in the test set. It is indicated from the figure that the deviation in the ground-truth and predicted values are not so large and that the predictive performance can be further improved by improving the training dataset in terms of quantity and representation. To quantitatively assess the performance of our model, we computed three correlation measures: the PLCC, the SROCC, and the KROCC. These metrics were calculated by comparing the predicted quality scores to the ground truth scores for both the test set of the datasets. The results for each of these metrics are presented in Table 1 & 2. This table provides a comprehensive summary of the model's performance, allowing for a clear comparison of its predictive accuracy and consistency across different evaluation metrics and datasets.

E. COMPARISON WITH EXISTING APPROACHES

The experimental results of the proposed approach for both datasets are presented in Table 1 & 2, which showcases the performance of the model. To evaluate the performance of the proposed approach in comparison to state-of-the-art methods, we selected 25 existing approaches and conducted a comprehensive comparison of their predictive performance.

The comparison is based on several key metrics, including SROCC, KROCC, and PLCC. These metrics measure the correlation between the predicted quality scores and the ground truth quality scores. The higher the correlation values, the stronger the relationship between the predicted and actual quality scores. In this comparison, we excluded Root Mean Square Error (RMSE) as a metric since we performed re-scaling of the Mean Opinion Scores (MOS) to a range of 0 to 1 and therefore it will not reflect the RMSE as an accurate measure for comparison. However, we included an important metric, namely, the Computation Time Ratio (CTR), which represents the ratio of prediction time to video playback duration. A lower ratio indicates a model with low computational complexity, and a value lower than 1 indicates that the model meets the real-time prediction requirement (i.e. the quality can be predicted before the playback duration of the video). Thus, we considered CTR as a metric to assess and discuss the performance of competitive models based on their time efficiency. It is worth noting that a quality prediction approaches that can provide inference within a CTR of less than 1 is considered suitable for real-time operations, and used this criterion to compare models.

TABLE 1. Performance assessment on "Waterloo sQoE III" dataset.

Sr.	Approach	SROCC	KROCC	PLCC	Time
1	Yin2015 [25]	0.146	0.093	0.323	0.007
2	SQI [30]	0.152	0.110	0.223	0.500
3	MoK2012 [26]	0.169	0.129	0.216	0.001
4	FTW [27]	0.184	0.134	0.323	0.001
5	Bentaleb20l6 [31]	0.198	0.139	0.341	0.500
6	Liu2012 [28]	0.253	0.172	0.242	0.001
7	Xue2014 [29]	0.341	0.225	0.308	0.003
8	VIIDEO [36]	0.395	0.265	0.490	8.050
9	NIQE [35]	0.402	0.273	0.449	1.030
10	PIQE [55]	0.423	0.281	0.435	1.130
11	BRISQUE [34]	0.496	0.347	0.469	1.940
12	TV-QoE [32]	0.507	0.357	0.467	0.520
13	KSQI [33]	0.529	0.388	0.527	0.510
14	ResNet50 [56]	0.571	0.411	0.564	0.380
15	UGC-VQA [38]	0.590	0.424	0.646	1.450
16	PIQI [11]	0.495	0.365	0.590	0.920
17	VIDEVAL [57]	0.686	0.465	0.652	0.730
18	DeepEns [1]	0.696	0.547	0.669	0.620
19	VSFA [40]	0.704	0.489	0.669	1.210
20	TLVQM [58]	0.726	0.493	0.689	1.100
21	V-BLIINDS [59]	0.739	0.546	0.724	45.63
22	FAST-VQA [41]	0.739	0.550	0.771	0.330
23	PVQ [39]	0.749	0.551	0.727	2.070
24	StarVQA [42]	0.757	0.536	0.706	4.100
25	TQP (Proposed)	0.782	0.713	0.721	0.230

TABLE 2. Performance assessment on "LIVE Netflix II" dataset.

Sr.	Approach	SROCC	KROCC	PLCC	Time
1	MoK2012 [26]	0.080	0.065	0.087	0.001
2	Yin2015 [25]	0.080	0.062	0.065	0.007
3	FTW [27]	0.080	0.086	0.065	0.001
4	VIIDEO [36]	0.284	0.189	0.323	1.276
5	BRISQUE [34]	0.315	0.218	0.201	5.624
6	ResNet50 [56]	0.428	0.300	0.432	0.307
7	Bentaleb20l6 [31]	0.445	0.298	0.453	0.456
8	PIQI [11]	0.503	0.382	0.624	0.862
10	Xue2014 [29]	0.583	0.412	0.496	0.003
11	UGC-VQA [38]	0.597	0.456	0.684	1.353
12	NIQE [35]	0.654	0.476	0.668	0.529
13	Liu2012 [28]	0.663	0.468	0.637	0.001
14	TV-QoE [32]	0.669	0.414	0.511	0.482
15	PIQE [55]	0.675	0.490	0.687	0.541
16	VIDEVAL [57]	0.698	0.491	0.687	0.679
17	DeepEns [1]	0.705	0.584	0.707	0.587
18	VSFA [40]	0.715	0.515	0.700	1.134
19	SQI [30]	0.735	0.530	0.633	0.458
20	KSQI [33]	0.739	0.549	0.732	0.462
21	TLVQM [58]	0.740	0.520	0.719	1.025
22	V-BLIINDS [59]	0.751	0.576	0.765	42.75
23	FAST-VQA [41]	0.751	0.586	0.807	0.309
24	PVQ [39]	0.753	0.582	0.767	1.940
25	StarVQA [42]	0.779	0.572	0.739	3.842
26	TQP (Proposed)	0.795	0.652	0.772	0.216

Table 1 presents the performance comparison of the proposed approach using the Waterloo sQoE-III dataset, while Table 2 showcases the performance comparison using the LIVE Netflix-II dataset. These tables provide a comprehensive analysis of the proposed approach's performance, comparing it to other state-of-the-art methods across various correlation measures and CTR. By conducting this extensive comparison, we gain insights into the effectiveness and efficiency of the proposed approach for VQA, establishing its competitive edge and suitability for real-time applications.

F. CROSS-DATASET EVALUATION

Cross-dataset evaluations were performed to demonstrate the robustness of the proposed TQP model for assessing video quality. This involved testing models on the LIVE Netflix II dataset that were trained on the Waterloo sQoE III dataset, and vice versa. Cross-dataset evaluation is crucial as it allows for an assessment of the model's generalization capability beyond the training dataset. The study demonstrates the model's ability to generalize and accurately predict video quality across diverse datasets and real-world scenarios by ensuring that it performs consistently well on datasets other than those on which it was trained. In order to compare cross-dataset evaluation, some existing models are trained and evaluated on the same scenario and are presented for comparison in Table 3.

These results indicate that the TQP model demonstrates strong generalization capabilities across datasets. Despite being trained on one dataset, it exhibits consistent and high-performance levels when tested on a different dataset. The similarity in performance metrics between the training and testing datasets suggests that the TQP model effectively learns and captures underlying video quality features that are transferable across datasets, rather than being overly tailored to dataset-specific characteristics. Therefore, the cross-dataset evaluation results validate the robustness and effectiveness of the TQP model in predicting video quality, underscoring its applicability and reliability across diverse datasets and real-world scenarios.

G. ABLATION STUDY

The ablation study results highlight the impact of using different backbone architectures with and without the proposed channel shift on the performance metrics SROCC, KROCC, and PLCC. These metrics are crucial for assessing the quality of video predictions in terms of ranking consistency, correlation, and linearity with ground truth scores. The key observations from the study are as follows:

- **Performance Improvement with Channel Shift:** The performance improvement with the application of channel shift is evident across all backbone architectures, showing significant enhancement in SROCC, KROCC, and PLCC metrics. This demonstrates the effectiveness of channel shift in improving model performance. Notably, the EfficientNet-B0 architecture exhibits the most substantial improvement across all three metrics. This confirms its superior effectiveness and efficiency when combined with channel shift, making it the standout performer among the tested architectures.
- **Backbone Architectures Without Channel Shift:** Among the backbone architectures without the application of channel shift, SqueezeNet achieves the lowest performance, with SROCC, KROCC, and PLCC values of 0.527, 0.392, and 0.512, respectively. MobileNet-V2 shows moderate performance, achieving scores of 0.608, 0.424, and 0.594. ResNet-50 performs slightly better than SqueezeNet but still has limited effectiveness, with scores of 0.571, 0.411, and 0.564. ShuffleNet is the lowest performer overall, with values of 0.506, 0.385, and 0.504. EfficientNet-B0 stands out as the best performer in this group, with scores of 0.623, 0.587, and 0.619, demonstrating its relatively superior performance even without channel shift.
- **Backbone Architectures With Channel Shift:** Among the backbone architectures with the application of channel shift, SqueezeNet shows notable improvement, with scores rising to 0.692, 0.665, and 0.681. MobileNet-V2 achieves higher scores of 0.768, 0.724, and 0.748, demonstrating significant enhancement. ResNet-50 also shows considerable improvement, with values increasing to 0.762, 0.702, and 0.746. ShuffleNet, while improved, still performs lower than the other architectures, with scores of 0.651, 0.610, and 0.634. EfficientNet-B0 achieves the highest performance, with scores of 0.782, 0.713, and 0.721, establishing it as the best-performing model with channel shift applied.
- **Proposed Model:** The EfficientNet-B0 architecture with channel shift stands out as the proposed model in this study, delivering the best results across all three metrics and demonstrating its superiority in effectively leveraging temporal features in video quality assessment. The ablation study clearly shows that incorporating the proposed channel shift method enhances the performance of all tested architectures. EfficientNet-B0 with channel shift emerges as the top-performing model, providing robust and reliable video quality predictions suitable for adaptive bitrate video streaming applications. This study underscores the importance of selecting an efficient backbone and integrating advanced techniques like channel shift to optimize performance.

H. DISCUSSION

Tables 1 and 2 provide a comprehensive analysis of the comparative performance of the proposed TQP approach on the Waterloo sQoS III dataset and the Live Netflix II dataset, respectively. The evaluation is based on important metrics such as PLCC, SROCC, KROCC, and computation Time ratio (representing the computational efficiency of the approach).

To assess the performance of video quality prediction models, a higher correlation coefficient is desirable, indicating

TABLE 3. Results of cross-dataset evaluation.

Method	Training Dataset	Testing Dataset	SROCC	KROCC	PLCC
PIQE [55]	Waterloo sQoE III	LIVE Netflix II	0.643	0.619	0.649
	LIVE-NFLX-II	Waterloo sQoE III	0.576	0.560	0.602
BMPRI [60]	Waterloo sQoE III	LIVE Netflix II	0.504	0.475	0.503
	LIVE-NFLX-II	Waterloo sQoE III	0.455	0.427	0.428
IL-NIQE [61]	Waterloo sQoE III	LIVE Netflix II	0.492	0.493	0.518
	LIVE-NFLX-II	Waterloo sQoE III	0.414	0.407	0.416
NIQE [35]	Waterloo sQoE III	LIVE Netflix II	0.483	0.477	0.478
	LIVE-NFLX-II	Waterloo sQoE III	0.458	0.454	0.457
PaQ-2-PiQ [62]	Waterloo sQoE III	LIVE Netflix II	0.427	0.408	0.417
	LIVE-NFLX-II	Waterloo sQoE III	0.415	0.391	0.441
DIIVINE [59]	Waterloo sQoE III	LIVE Netflix II	0.743	0.698	0.726
	LIVE-NFLX-II	Waterloo sQoE III	0.715	0.716	0.730
BRISQUE [34]	Waterloo sQoE III	LIVE Netflix II	0.690	0.681	0.685
	LIVE-NFLX-II	Waterloo sQoE III	0.687	0.667	0.705
PIQI [11]	Waterloo sQoE III	LIVE Netflix II	0.721	0.709	0.717
	LIVE-NFLX-II	Waterloo sQoE III	0.651	0.629	0.671
CORNIA [63]	Waterloo sQoE III	LIVE Netflix II	0.712	0.691	0.732
	LIVE-NFLX-II	Waterloo sQoE III	0.656	0.624	0.659
NBIQA [64]	Waterloo sQoE III	LIVE Netflix II	0.583	0.581	0.584
	LIVE-NFLX-II	Waterloo sQoE III	0.495	0.472	0.509
HOSA [65]	Waterloo sQoE III	LIVE Netflix II	0.672	0.617	0.660
	LIVE-NFLX-II	Waterloo sQoE III	0.606	0.569	0.601
DeepEns [1]	Waterloo sQoE III	LIVE Netflix II	0.721	0.728	0.748
	LIVE-NFLX-II	Waterloo sQoE III	0.721	0.685	0.738
FRIQUEE [66]	Waterloo sQoE III	LIVE Netflix II	0.617	0.613	0.622
	LIVE-NFLX-II	Waterloo sQoE III	0.545	0.538	0.557
TQP (Proposed)	LIVE Netflix II	Waterloo sQoE III	0.745	0.738	0.75
	LIVE-NFLX-II	LIVE Netflix II	0.763	0.754	0.76

TABLE 4. Results of ablation study.

Backbone	Depth	∼Parameters	∼Size	Channel Shift	SROCC	KROCC	PLCC
SqueezeNet	18 layers	1.24 million	5.2 MB	Without Channel Shift	0.527	0.392	0.512
MobileNet-V2	53 layers	3.5 million	13 MB		0.608	0.424	0.594
ResNet-50	50 layers	25.6 million	96 MB		0.571	0.411	0.564
ShuffleNet	50 layers	1.4 million	5.4 MB		0.506	0.385	0.504
EfficientNet-B0	82 layers	5.3 million	20 MB		0.623	0.587	0.619
SqueezeNet	18 layers	1.24 million	5.2 MB	With Channel Shift	0.692	0.665	0.681
MobileNet-V2	53 layers	3.5 million	13 MB		0.768	0.724	0.748
ResNet-50	50 layers	25.6 million	96 MB		0.762	0.702	0.746
ShuffleNet	50 layers	1.4 million	5.4 MB		0.651	0.61	0.634
EfficientNet-B0	82 layers	5.3 million	20 MB		0.782	0.713	0.721

a stronger relationship between predicted and ground truth quality scores. Additionally, a low Time value (less than 1) is preferred for real-time performance, as it signifies that the inference time is shorter than the playback duration.

In the evaluation, the proposed TQP approach is compared against twenty-four existing approaches. The methods are ranked in ascending order based on their SROCC values obtained from both datasets. The results demonstrate that the proposed TQP approach exhibits the best prediction performance in terms of both PLCC and SROCC, while maintaining real-time computational efficiency. Its inference time accounts for only 21.6% of the video's runtime, making it highly suitable for real-world implementation.

Among the evaluated approaches, StarVQA [42] stands out as the second-best performing technique based on SROCC. However, its KROCC and PLCC scores are comparatively lower than those of the other approaches. Furthermore, StarVQA's [42] computational efficiency is non-real-time, requiring 4.1 times the playback duration. In the LIVE Netflix II dataset, FastVQA [41] emerges with the highest KROCC value and the second-highest PLCC value. Additionally, FastVQA [41] achieves real-time processing efficiency, utilizing approximately 30.9% of the video's playback duration. PVQ [39] is another competitive strategy, exhibiting excellent predictive performance, albeit with slightly lower computational efficiency than real-time operation.

Moreover, it is worth mentioning that MoK2012 [26], FTW [27], and Liu et al. [28] are the fastest approaches for video quality prediction, with inference times as low as 0.1% of the video playback time. However, these approaches demonstrate relatively low predictive performance, with Liu et al. [28] being the best performer among the extremely fast techniques. In the case of the Live Netflix II dataset, Liu et al. [28] achieves an SROCC of 0.663, KROCC of 0.468, and PLCC of 0.637, while making predictions in

only 0.1% of the video playback time. Thus, in scenarios with limited processing capabilities, Liu et al. [28] may be a suitable choice.

As a closing note, when considering the trade-off between predictive performance and computational complexity, the proposed TQP approach and FastVQA [41] emerge as the most favorable choices. These approaches strike a balance between accuracy and efficiency, making them well-suited for practical deployment in video quality assessment applications.

V. CONCLUSION

The importance of efficient rate adaptation to satisfy network requirements has dramatically increased with the widespread use of video streaming services and the shared nature of the internet. Nonetheless, excessively optimistic rate adaptation can result in poor Quality of Experience (QoE) for the user. The objective evaluation of video quality is essential for enhancing QoE and facilitating optimal rate adaptation.

This research addressed the challenge of video quality assessment for rate adaptive videos. Existing deep learning-based video quality evaluation models often prioritize high prediction performance at the expense of computational efficiency, making them unsuitable for real-time deployment. To overcome this limitation, we developed a reliable and cost-effective quality assessment system that leverages temporal learning using 2D CNN-based quality assessment models with channel shifting. The proposed approach introduces channel shifting, which transfers 1/8th of the features from a frame to its subsequent frame. This technique effectively captures temporal information, enhancing the quality assessment process. The experimental results demonstrate that our approach achieves state-of-the-art performance in terms of correlation with human judgment, indicating its superior predictive capabilities. Moreover, one of the key strengths of our approach lies in its computational efficiency, enabling real-time operation. The proposed system is capable of performing quality prediction in just 23% of the video's duration, making it highly practical for real-world applications.

In conclusion, our research offers a legitimate alternative for objective quality evaluation in video streaming applications. The proposed method establishes a balance between precision and computational efficiency, resulting in enhanced QoE and optimal rate adaptation. Our system contributes to improving user satisfaction and the overall streaming experience by providing a dependable and cost-effective video quality evaluation.

A. FUTURE RESEARCH DIRECTIONS

While this study has made significant contributions to video quality assessment for rate adaptive videos, there are several avenues for future research that could further advance the field. Some potential directions for future investigations include:

- Integrate quality assessment with dynamic rate adaptation for real-time adaptive streaming, optimizing user QoE by dynamically adjusting video quality based on assessed scores.
- Incorporate subjective aspects of video quality by considering user preferences and perceptions, personalizing the assessment system for tailored QoE.
- Evaluate system performance on diverse video content, enhancing robustness across genres, resolutions, and encoding formats.
- Conducting field trials and user studies to evaluate the proposed system in real-world scenarios and collect user feedback. Assessing the system's performance under various network conditions and user environments to validate its effectiveness and practicality.
- Contribute to standardized evaluation protocols and benchmark datasets for video quality assessment in rate adaptive videos, promoting reproducibility and facilitating comparisons.

AUTHORS CONTRIBUTION

All the authors have contributed equally to the study.

DECLARATION OF COMPETING INTEREST

The authors declare the following financial interests/personal relationships, which may be considered potential competing interests: Muhammad Azeem Aslam reports that the administration of Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China, provided support. Muhammad Azeem Aslam writes a connection with Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun 130033, China. Muhammad Azeem Aslam has confirmed no conflict of interest for this study.

REFERENCES

[1] N. Ahmed, H. M. Shahzad Asif, A. R. Bhatti, and A. Khan, "Deep ensembling for perceptual image quality assessment," *Soft Comput.*, vol. 26, no. 16, pp. 7601–7622, Aug. 2022.
[2] J. Li, L. Han, C. Zhang, Q. Li, and Z. Liu, "Spherical convolution empowered viewport prediction in 360 video multicast with limited FoV feedback," *ACM Trans. Multimedia Comput., Commun., Appl.*, vol. 19, no. 1, pp. 1–23, Jan. 2023.
[3] S. Pan, G. J. W. Xu, K. Guo, S. H. Park, and H. Ding, "Video-based engagement estimation of game streamers: An interpretable multimodal neural network approach," *IEEE Trans. Games*, pp. 1–12, 2023.
[4] E. S. Gama, L. O. N. De Araújo, R. Immich, and L. F. Bittencourt, "Video streaming analysis in multi-tier edge-cloud networks," in *Proc. 8th Int. Conf. Future Internet Things Cloud*, Aug. 2021, pp. 19–25.
[5] J. Li, C. Zhang, Z. Liu, R. Hong, and H. Hu, "Optimal volumetric video streaming with hybrid saliency based tiling," *IEEE Trans. Multimedia*, 2022.
[6] F. Lozano, M.-C. Aguayo-Torres, G. Gómez, C. Cárdenas, and J. Baños, "Network traffic analysis and qoe evaluation for video progressive download service: Netflix," in *Proc. Int. Conf. Wired Wireless Internet Commun.*, 2015, pp. 239–246.
[7] M. Azeem Aslam, X. Wei, N. Ahmed, G. Saleem, T. Amin, and H. Caixue, "VRL-IQA: Visual representation learning for image quality assessment," *IEEE Access*, vol. 12, pp. 2458–2473, 2024.
[8] H. Khalid and N. Ahmed, "Blind image quality assessment using multi-stream architecture with spatial and channel attention," 2023, *arXiv:2307.09857*.

[9] Q. Zheng, Z. Tu, P. C. Madhusudana, X. Zeng, A. C. Bovik, and Y. Fan, "FAVER: Blind quality prediction of variable frame rate videos," *Signal Process., Image Commun.*, vol. 122, Mar. 2024, Art. no. 117101.

[10] Z. Cui, H. Sheng, D. Yang, S. Wang, R. Chen, and W. Ke, "Light field depth estimation for non-lambertian objects via adaptive cross operator," *IEEE Trans. Circuits Syst. Video Technol.*, 2023.

[11] N. Ahmed, H. M. S. Asif, and H. Khalid, "PIQI: Perceptual image quality index based on ensemble of Gaussian process regression," *Multimedia Tools Appl.*, vol. 80, no. 10, pp. 15677–15700, Apr. 2021.

[12] Q. Liu, H. Yuan, R. Hamzaoui, H. Su, J. Hou, and H. Yang, "Reduced reference perceptual quality model with application to rate control for video-based point cloud compression," *IEEE Trans. Image Process.*, vol. 30, pp. 6623–6636, 2021.

[13] J. Yan, L. Wu, W. Jiang, C. Liu, and F. Shen, "Revisiting the robustness of spatio-temporal modeling in video quality assessment," *Displays*, vol. 81, Jan. 2024, Art. no. 102585.

[14] Y. Fang, Z. Li, J. Yan, X. Sui, and H. Liu, "Study of spatio-temporal modeling in video quality assessment," *IEEE Trans. Image Process.*, vol. 32, pp. 2693–2702, 2023.

[15] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE Trans. Image Process.*, vol. 19, no. 2, pp. 335–350, Feb. 2010.

[16] N. Ahmed and S. Asif, "BIQ2021: A large-scale blind image quality assessment database," 2022, *arXiv:2202.03879*.

[17] N. Ahmed and H. M. S. Asif, "Ensembling convolutional neural networks for perceptual image quality assessment," in *Proc. 13th Int. Conf. Math., Actuarial Sci., Comput. Sci. Statist. (MACS)*, Dec. 2019, pp. 1–5.

[18] N. Ahmed and H. M. S. Asif, "Perceptual quality assessment of digital images using deep features," *Comput. Informat.*, vol. 39, no. 3, pp. 385–409, 2020.

[19] N. Ahmad, H. M. S. Asif, G. Saleem, M. U. Younus, S. Anwar, and M. R. Anjum, "Leaf image-based plant disease identification using color and texture features," *Wireless Pers. Commun.*, vol. 121, no. 2, pp. 1139–1168, Nov. 2021.

[20] G. Saleem, M. Akhtar, N. Ahmed, and W. S. Qureshi, "Automated analysis of visual leaf shape features for plant classification," *Comput. Electron. Agricult.*, vol. 157, pp. 270–280, Feb. 2019.

[21] S. Nawaz, A. Calefati, N. Ahmed, and I. Gallo, "Hand written characters recognition via deep metric learning," in *Proc. 13th IAPR Int. Workshop Document Anal. Syst. (DAS)*, Apr. 2018, pp. 417–422.

[22] S. Alamgeer, M. Irshad, and M. C. Q. Farias, "CNN-based no-reference video quality assessment method using a spatiotemporal saliency patch selection procedure," *J. Electron. Imag.*, vol. 30, no. 6, Nov. 2021, Art. no. 063001.

[23] Y. Zheng, P. Liu, L. Qian, S. Qin, X. Liu, Y. Ma, and G. Cheng, "Recognition and depth estimation of ships based on binocular stereo vision," *J. Mar. Sci. Eng.*, vol. 10, no. 8, p. 1153, Aug. 2022.

[24] W. Wu, H. Zhu, S. Yu, and J. Shi, "Stereo matching with fusing adaptive support weights," *IEEE Access*, vol. 7, pp. 61960–61974, 2019.

[25] X. Yin, A. Jindal, V. Sekar, and B. Sinopoli, "A control-theoretic approach for dynamic adaptive video streaming over HTTP," in *Proc. ACM Conf. Special Interest Group Data Commun.*, Aug. 2015, pp. 325–338.

[26] R. K. P. Mok, X. Luo, E. W. W. Chan, and R. K. C. Chang, "QDASH: A QoE-aware DASH system," in *Proc. 3rd Multimedia Syst. Conf.*, Feb. 2012, pp. 11–22.

[27] T. Hoßfeld, R. Schatz, E. Biersack, and L. Plissonneau, "Internet video delivery in YouTube: From traffic measurements to quality of experience," in *Data Traffic Monitoring and Analysis*, 2013, pp. 264–301.

[28] X. Liu, F. Dobrian, H. Milner, J. Jiang, V. Sekar, I. Stoica, and H. Zhang, "A case for a coordinated Internet video control plane," in *Proc. ACM SIGCOMM Conf. Appl., Technol., Architectures, Protocols Comput. Commun.*, Aug. 2012, pp. 359–370.

[29] J. Xue, D.-Q. Zhang, H. Yu, and C. Wen Chen, "Assessing quality of experience for adaptive HTTP video streaming," in *Proc. IEEE Int. Conf. Multimedia Expo. Workshops (ICMEW)*, Jul. 2014, pp. 1–6.

[30] Z. Duanmu, K. Zeng, K. Ma, A. Rehman, and Z. Wang, "A quality-of-experience index for streaming video," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 154–166, Feb. 2017.

[31] A. Bentaleb, A. C. Begen, and R. Zimmermann, "SDNDASH: Improving QoE of HTTP adaptive streaming using software defined networking," in *Proc. 24th ACM Int. Conf. Multimedia*, Oct. 2016, pp. 1296–1305.

[32] D. Ghadiyaram, J. Pan, and A. C. Bovik, "Learning a continuous-time streaming video QoE model," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2257–2271, May 2018.

[33] Z. Duanmu, W. Liu, D. Chen, Z. Li, Z. Wang, Y. Wang, and W. Gao, "A knowledge-driven quality-of-experience model for adaptive streaming videos," 2019, *arXiv:1911.07944*.

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

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

[36] A. Mittal, M. A. Saad, and A. C. Bovik, "A completely blind video integrity Oracle," *IEEE Trans. Image Process.*, vol. 25, no. 1, pp. 289–300, Jan. 2016.

[37] M. Tan and Q. Le, "EfficientNet: Rethinking model scaling for convolutional neural networks," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 6105–6114.

[38] W. Sun, X. Min, W. Lu, and G. Zhai, "A deep learning based no-reference quality assessment model for UGC videos," in *Proc. 30th ACM Int. Conf. Multimedia*, Oct. 2022, pp. 856–865.

[39] Z. Ying, M. Mandal, D. Ghadiyaram, and A. Bovik, "Patch-VQ: 'patching up' the video quality problem," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2021, pp. 14014–14024.

[40] D. Li, T. Jiang, and M. Jiang, "Quality assessment of in-the-wild videos," in *Proc. 27th ACM Int. Conf. Multimedia*, Oct. 2019, pp. 2351–2359.

[41] H. Wu, C. Chen, J. Hou, L. Liao, A. Wang, W. Sun, Q. Yan, and W. Lin, "FAST-VQA: Efficient end-to-end video quality assessment with fragment sampling," in *Computer Vision–ECCV*. Cham, Switzerland: Springer, 2022, pp. 538–554.

[42] F. Xing, Y.-G. Wang, H. Wang, L. Li, and G. Zhu, "StarVQA: Space-time attention for video quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2022, pp. 2326–2330.

[43] M. Xu, J. Chen, H. Wang, S. Liu, G. Li, and Z. Bai, "C3DVQA: Full-reference video quality assessment with 3D convolutional neural network," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 4447–4451.

[44] G. Varol, I. Laptev, and C. Schmid, "Long-term temporal convolutions for action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 40, no. 6, pp. 1510–1517, Jun. 2018.

[45] J. Yang, Y. Zhu, C. Ma, W. Lu, and Q. Meng, "Stereoscopic video quality assessment based on 3D convolutional neural networks," *Neurocomputing*, vol. 309, pp. 83–93, Oct. 2018.

[46] J. You and J. Korhonen, "Deep neural networks for no-reference video quality assessment," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 2349–2353.

[47] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 4489–4497.

[48] J. Carreira and A. Zisserman, "Quo vadis, action recognition? A new model and the kinetics dataset," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 4724–4733.

[49] O. Köpüklü, N. Kose, A. Gunduz, and G. Rigoll, "Resource efficient 3D convolutional neural networks," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. Workshop (ICCVW)*, Oct. 2019, pp. 1910–1919.

[50] G. Saleem, U. I. Bajwa, R. Hammad Raza, F. H. Alqahtani, A. Tolba, and F. Xia, "Efficient anomaly recognition using surveillance videos," *PeerJ Comput. Sci.*, vol. 8, Oct. 2022, Art. no. e1117.

[51] J. Lin, C. Gan, and S. Han, "TSM: Temporal shift module for efficient video understanding," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 7082–7092.

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

[53] C. G. Bampis, Z. Li, I. Katsavounidis, T.-Y. Huang, C. Ekanadham, and A. C. Bovik, "Towards perceptually optimized end-to-end adaptive video streaming," 2018, *arXiv:1808.03898*.

[54] Z. Duanmu, A. Rehman, and Z. Wang, "A quality-of-experience database for adaptive video streaming," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 474–487, Jun. 2018.

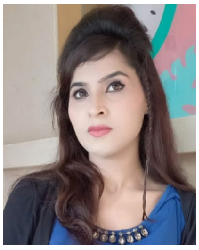
[55] N. Venkatanath, D. Praneeth, M. C. Bh, S. S. Channappayya, and S. S. Medasani, "Blind image quality evaluation using perception based features," in *Proc. 21st Nat. Conf. Commun. (NCC)*, Feb. 2015, pp. 1–6.

[56] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.
[57] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, and A. C. Bovik, "UGC-VQA: Benchmarking blind video quality assessment for user generated content," *IEEE Trans. Image Process.*, vol. 30, pp. 4449–4464, 2021.
[58] J. Korhonen, "Two-level approach for no-reference consumer video quality assessment," *IEEE Trans. Image Process.*, vol. 28, no. 12, pp. 5923–5938, Dec. 2019.
[59] M. A. Saad, A. C. Bovik, and C. Charrier, "Blind prediction of natural video quality," *IEEE Trans. Image Process.*, vol. 23, no. 3, pp. 1352–1365, Mar. 2014.
[60] X. Min, G. Zhai, K. Gu, Y. Liu, and X. Yang, "Blind image quality estimation via distortion aggravation," *IEEE Trans. Broadcast.*, vol. 64, no. 2, pp. 508–517, Jun. 2018.
[61] Y. Liu, K. Gu, Y. Zhang, X. Li, G. Zhai, D. Zhao, and W. Gao, "Unsupervised blind image quality evaluation via statistical measurements of structure, naturalness, and perception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 30, no. 4, pp. 929–943, Apr. 2020.
[62] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 3572–3582.
[63] P. Ye, J. Kumar, L. Kang, and D. Doermann, "Unsupervised feature learning framework for no-reference image quality assessment," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 1098–1105.
[64] F.-Z. Ou, Y.-G. Wang, and G. Zhu, "A novel blind image quality assessment method based on refined natural scene statistics," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2019, pp. 1004–1008.
[65] J. Xu, P. Ye, Q. Li, H. Du, Y. Liu, and D. Doermann, "Blind image quality assessment based on high order statistics aggregation," *IEEE Trans. Image Process.*, vol. 25, no. 9, pp. 4444–4457, Sep. 2016.
[66] D. Ghadiyaram and A. C. Bovik, "Feature maps driven no-reference image quality prediction of authentically distorted images," in *Human Vision and Electronic Imaging*, vol. 9394. Bellingham, WA, USA: SPIE, 2015, pp. 158–171.

MUHAMMAD AZEEM ASLAM received the Ph.D. degree from Northwestern Polytechnical University, Xi'an, China. He is currently with Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, and the School of Information Engineering, Xi'an Eurasia University, Shanxi, China. His research interests include computer vision and machine learning.

XU WEI received the B.S. degree in mechanical and electronic engineering from Jilin University, Changchun, China, in 2003, and the Ph.D. degree in mechanical and electronic engineering from Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, in 2008. Since 2008, he has been with Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences. He is currently a Research Fellow and the Ph.D. Supervisor. His current research interests include the integration technology of satellites and payloads and the highly reliable electronic systems for aerospace.

NISAR AHMED received the master's and Ph.D. degrees in computer engineering from the University of Engineering and Technology at Lahore, Lahore, Pakistan. With over 13 years of dedicated research and professional experience, he has made significant contributions to the fields of digital image processing, computer vision, machine learning, and data science. His current research interests include pattern recognition, computer vision, and digital image and video processing.



GULSHAN SALEEM received the master's degree in software engineering from the College of E&ME, National University of Science and Technology, Rawalpindi, Pakistan, in 2016. She is currently pursuing the Ph.D. degree in computer science with COMSATS University Islamabad, Lahore Campus, Pakistan. She is also a Lecturer with the Department of Computer Science, Lahore Garrison University, Lahore, Pakistan. Her research interests include computer vision, machine learning, and digital image processing. She is passionate about exploring innovative solutions at the intersection of these fields to address contemporary challenges in computer science and technology.

ZHU SHUANGTONG received the master's degree in engineering. She is currently with Changchun Institute of Fine Mechanics and Physics, Chinese Academy of Sciences, China. Her current research interests include remote sensing, image processing, and software engineering.

YIMEI XU received the bachelor's and master's degrees in computer sciences. She is currently with the School of Information Engineering, Xi'an Eurasia University. Her current research interests include computer vision and machine learning.

HU HONGFEI received the bachelor's and master's degrees in computer engineering. He is currently with the School of Information Engineering, Xi'an Eurasia University. His current research interests include computer vision and image processing.