RESEARCH ARTICLE

3D Convolutional Neural Network-Based Multi-Parameter Video Quality Assessment Model on Cloud Platforms

Xue Li[1,*] **and Jiali Qiu**[2]

[1] Xidian University, Xi'an 710126, China
[2] Xi'an Jiaotong University,Xi'an 710049, China

Abstract

In light of the rapid advancements in big data and artificial intelligence technologies, the trend of uploading local files to cloud servers to mitigate local storage limitations is growing. However, the surge of duplicate files, especially images and videos, results in significant network bandwidth wastage and complicates server management. To tackle these issues, we have developed a multi-parameter video quality assessment model utilizing a 3D convolutional neural network within a video deduplication framework. Our method, inspired by the analytic hierarchy process, thoroughly evaluates the effects of packet loss rate, codec, frame rate, bit rate, and resolution on video quality. The model employs a two-stream 3D convolutional neural network to integrate spatial and temporal streams for capturing video distortion details, with a coding layer configured to remove redundant distortion information. We validated our approach using the LIVE and CSIQ datasets, comparing its performance against the V-BLIINDS and VIDEO schemes across different packet loss rates. Furthermore, we simulated the client-server interaction using a subset of the dataset and assessed the scheme's time efficiency. Our results indicate that the proposed scheme offers a highly efficient solution for video quality assessment.

Keywords: Video quality assessment, 3D CNN, Packet loss rate, SRCC, PLCC.

Citation
Li, X., & Qiu, J. (2024). 3D Convolutional Neural Network-Based Multi-Parameter Video Quality Assessment Model on Cloud Platforms. IECE Transactions on Internet of Things, 2(1), 8–19.

1 Introduction

The advent of big data, coupled with the proliferation of portable shooting devices like digital cameras and smartphones, has significantly amplified the demand for data storage solutions. Due to the constraints of local storage space, many users prefer to store high-definition pictures and videos on cloud servers. However, as more users repeatedly upload multimedia files, this practice not only results in substantial network bandwidth wastage but also leads to significant data redundancy, complicating the daily management of cloud storage systems.

According to the International Data Corporation, the volume of digital data reached 44ZB in 2020, with approximately 75% of this data being duplicates. Furthermore, data redundancy on cloud servers used for backup and storage exceeds 90%. Consequently, detecting and deleting duplicate multimedia files has become a critical task.

Data deduplication technology [1] ensures that only a single copy of each file is maintained on the server. Users who store similar files are provided with a

Academic Editor:
Jinchao Chen

Submitted: 17 November 2023
Accepted: 03 January 2024
Published: 14 January 2024

Vol. 2, **No.** 1, 2024.
10.62762/TIOT.2024.369369

***Corresponding author:**
✉ Xue Li
Lxue6632@protonmail.com

link to access the existing copy, thereby eliminating the need for redundant storage. If the file already exists on the server, users are not required to upload another copy. Cloud service [24–27] providers rely on deduplication technology to eliminate duplicate data, thereby reducing both bandwidth and storage requirements.

A comprehensive video deduplication scheme encompasses video copy detection, video ownership authentication, and video quality assessment. Video copy detection involves comparing client-side videos with those on the server to identify similar content. Video ownership authentication ensures that users retain ownership of their videos and can recover stored files from the server. Video quality assessment compares the quality of similar videos on both the client and server sides, retaining the higher quality video on the server and deleting the lower quality duplicates. Video deduplication can be categorized into server-based and client-based methods. Server-based deduplication involves uploading data from the client to the server and then deleting duplicates, which requires substantial upload bandwidth. Conversely, client-based deduplication deletes duplicate data before uploading, thereby saving significant upload bandwidth and server storage space [2]. As a result, client-based deduplication is the preferred method.

When users upload videos to the server, the User Datagram Protocol (UDP), a connectionless and unreliable transmission protocol, is often employed. This can result in packet loss during transmission, or issues with the original video such as low resolution and unsmooth playback. Therefore, it is essential to evaluate both the original quality of the video and the transmission process to assess the degree of video distortion.

Current research on video quality assessment in cloud environments is still in its early stages. To address this practical issue, we propose a method based on a 3D convolutional neural network. This method evaluates the impact of packet loss rate, codec, bit rate, frame rate, and resolution by extracting two-stream video distortion features from both spatial and temporal flows. The coding layer is utilized to reduce the redundancy of distortion information. Verification using the LIVE and CSIQ datasets indicates that the proposed scheme is highly efficient in video quality assessment. Additionally, our method offers significant improvements in accurately identifying and eliminating duplicate video content, thereby optimizing storage and bandwidth usage on cloud servers.

2 Related Works

The substantial storage requirements of videos, coupled with their higher network throughput demands compared to other multimedia content, make the elimination of duplicate videos on cloud servers imperative. To accurately evaluate video quality, it is essential to consider multiple factors comprehensively, such as packet loss during transmission, image clarity, and playback smoothness [13].

Researchers have proposed a variety of methods for video quality evaluation. Video Quality Assessment (VQA) algorithms [9, 10] are typically classified into three categories: full-reference (FR), reduced-reference (RR), and no-reference (NR). FR algorithms necessitate complete information from both the cloud and client videos for direct comparison, ensuring that the client's reference video is free from distortion. The quality is then assessed by measuring the differences between the reference and target videos. RR algorithms, however, use partial information from the reference video to make comparisons based on specific features, utilizing metrics such as MSE and PSNR. NR algorithms evaluate video quality without a reference video, relying instead on the intrinsic properties of the video, such as resolution and color.

In recent years, the remarkable success of deep convolutional neural networks in video feature extraction has led researchers to favor 3D convolutional neural networks for VQA. For instance, Li et al. [15] proposed an NR-VQA method utilizing a 3D shearlet transform and CNN to effectively capture anisotropic features of videos. Similarly, Yao et al. [3] introduced a bitrate-based metric that combines visual perception with robust generalization capabilities. Valderrama et al. [4] trained CNNs based on properties such as group of pictures (GOP) lengths and prioritization policies (BestEffort and DiffServ), although this approach was limited to low-resolution images and specific packet loss scenarios. Søgaard et al. [5] developed a regression function for video quality calculation, achieving correlation coefficients between 0.7 and 0.9 for dynamic and static video content. However, this method relied on image evaluation metrics, which can be significantly impacted when spatial video information is compromised.

Thus, selecting appropriate video quality evaluation

indicators is crucial, incorporating both spatial and temporal quality assessments and considering packet loss rates. Furthermore, Li et al. [22] proposed a unified NR-VQA framework with a mixed datasets training strategy for in-the-wild videos, based on VSFA, employing losses from monotonicity and linearity to handle mixed data training. Qian et al. [20] developed an innovative MIL-based model, VQA-MIL, which dynamically adjusts weights using a block-wise attention module and enhances video bag features with an MI Pooling layer. These advancements offer significant insights for VQA.

In our approach, we have developed a video quality assessment model based on a 3D convolutional neural network. This model captures video distortion information across both spatial and temporal dimensions, focusing on evaluating the impact of packet loss rate and codec on video quality. We employ Spearman's Rank Order Correlation Coefficient (SRCC) and Pearson's Linear Correlation Coefficient (PLCC) to evaluate the model's performance using the LIVE and CSIQ datasets. Our results demonstrate significant advantages in terms of video quality assessment efficiency and time cost.

Moreover, our method incorporates advanced techniques such as dropout and k-fold cross-validation to prevent overfitting, ensuring that the model generalizes well across diverse datasets. By leveraging the strengths of 3D convolutional neural networks, our approach not only improves the accuracy of video quality assessments but also enhances the robustness of the evaluations under various network conditions. The ability to effectively assess and manage video quality in cloud storage systems is crucial for optimizing storage and bandwidth usage, making our proposed scheme a valuable contribution to the field of video quality assessment.

3 Proposed Method

In this section, we elaborate on the unsupervised quality assessment model based on a 3D convolutional neural network from two perspectives: the scheme framework and the algorithm details.

3.1 Scheme Framework

The scheme framework consists of two parts: the first video uploader and subsequent video uploaders. The detailed process is illustrated in the following table.

Video Quality Assessment for the Initial Uploader

Client		Server
	Request to upload the quality score of video ←	
$quality(V) \rightarrow V_Q$		
	Upload the quality score parameter set $\{V, V_Q\}$ →	
		Save $\{V, V_Q\}$

(1) Following similarity detection and ownership verification, if no similar video exists on the server, the user is identified as the initial uploader. The server then requests the client to provide the video's quality score parameters.
(2) The client computes the quality score parameter $quality(V) \rightarrow V_Q$ and uploads $\{V, V_Q\}$ to the server.
(3) The server saves $\{V, V_Q\}$ in the database.

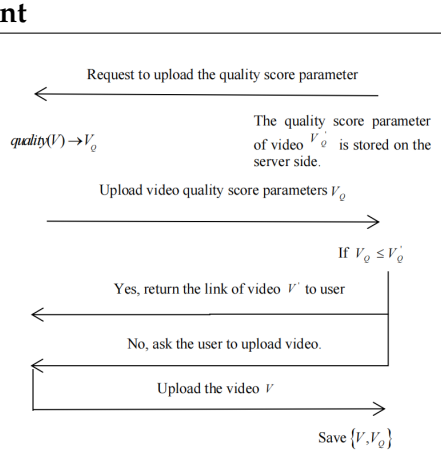
3.2 Algorithm details

In this section, we present a multi-parameter video quality assessment model tailored for cloud environments, utilizing a 3D convolutional neural network (CNN). This model is designed to calculate perceptual distortion values by analyzing temporal changes within video sequences. Recognizing that packet loss can significantly impact the measurement of video distortion, our approach incorporates weighted spatial entropy differences of various influencing factors—including packet loss, codec, bit rate, frame rate, and resolution—to more precisely capture video frame distortions. The detailed implementation is as follows:

Firstly, the model leverages a 3D CNN to analyze the video data across three dimensions: height, width, and time. This allows the network to better understand the temporal dynamics of video sequences, which are crucial for accurately assessing video quality. By extending the traditional 2D CNN framework to include the temporal dimension, our model can detect subtle changes and distortions that occur over time, providing a more comprehensive analysis.

Secondly, given the substantial effect of packet loss on video quality, our model assigns different weights to spatial entropy differences, which reflect the varying impact of each influencing factor on video distortion. This weighted approach enables the model to prioritize the most significant factors, such as packet loss, codec, bit rate, frame rate, and resolution, ensuring a more

Video Quality Assessment for Subsequent Uploaders



(1) After performing similarity detection and ownership verification, if the server already has a video with a quality score parameter V'_Q similar to video V', the client is requested to upload the quality score parameter of their video.
(2) Upon agreeing to the request, the client calculates the quality score parameter $quality(V) \rightarrow V_Q$ and uploads V_Q to the server.
(3) The server compares the quality parameters. If $V_Q \leq V'_Q$, indicating that the server-side video is of higher quality, the server provides the user with a link to the video and deletes the client's video to prevent duplication.
(4) If $V_Q > V'_Q$, indicating that the client-side video is of higher quality, the server requests the client to upload the complete video and quality parameters. The server then deletes the similar server-side video, freeing storage space, and stores $\{V, V_Q\}$ in the server database.

accurate assessment of video quality.

Thirdly, our model utilizes advanced preprocessing techniques to enhance the quality of the input data. This includes normalizing the video frames and applying data augmentation methods to increase the diversity of the training set. These steps are crucial for improving the robustness of the model and ensuring it performs well across a wide range of video content and network conditions.

Moreover, to prevent overfitting and improve the generalization of the model, we incorporate techniques such as dropout and k-fold cross-validation. Dropout helps in reducing the risk of overfitting by randomly omitting certain neurons during training, while k-fold cross-validation ensures that the model is validated on different subsets of the data, providing a more reliable measure of its performance.

Lastly, we evaluate the effectiveness of our proposed model using established metrics like Spearman's Rank Order Correlation Coefficient (SRCC) and Pearson's Linear Correlation Coefficient (PLCC). These metrics allow us to quantify the correlation between the predicted video quality scores and human subjective assessments, ensuring that our model provides results that are aligned with human perceptions of video quality.

By integrating these advanced methodologies, our multi-parameter video quality assessment model offers a sophisticated tool for evaluating video quality in cloud environments. This approach not only enhances the accuracy of video quality assessments but also provides a robust framework for managing video content on cloud servers, ultimately optimizing storage and bandwidth usage.

Algorithm 1: Maximum eigenvalue and eigenvector generation algorithm

```
function [c,e] = bll(a)
a=input('input the matrix')
vec=sum(a); %Start to normalize the matrix by
  column
[m,n]=size(a);
b=repmat(vec,m,1);
h=a./b; %Get the matrix normalized by column
c=[mean(h(1,:));mean(h(2,:));mean(h(3,:));
mean(h(4,:));mean(h(5,:))]
d=a*c;
e=(d(1)/c(1)+d(2)/c(2)+d(3)/c(3)+d(4)/c(4)
+d(5)/c(5))/5
c1=(e-5)/4
cr=c1/0.9 %Get the cr value
end
```

3.2.1 *Multivariate assessment*

We utilize a method akin to the Analytic Hierarchy Process (AHP) to assess the influence of packet loss rate, codec, bit rate, frame rate, and resolution on video quality. Among these, packet loss rate and codec are particularly critical factors impacting video quality [7]. Packet loss occurs when data packets encounter bit errors during transmission or when transmission delays exceed a set threshold, resulting in multiple data packets failing to reach their destination, thereby severely affecting video smoothness. Codecs

often degrade video quality because the compression process sacrifices video fidelity, as seen with codecs such as MPEG-2 and H.264. Additionally, bit rate, frame rate, and resolution are key determinants of video quality. Bit rate refers to the number of data bits transmitted per unit time; higher bit rates result in clearer videos. Frame rate denotes the number of frames transmitted per unit time; higher frame rates yield smoother videos. Resolution indicates the size of the video frame; higher resolutions produce larger video frames.

We define optimal video quality as the target layer and consider packet loss rate, codec, bit rate, frame rate, and resolution as criteria layers. The impact of these five factors on the target layer is then compared [11].

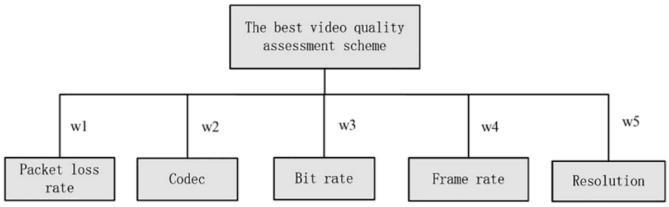


Figure 1. Hierarchy of the video quality assessment scheme.

A paired comparison matrix is constructed based on the degree of influence of each pair of factors on the target layer. Table 1 illustrates the meaning of the 1-9 comparison scale used in this method, where $\{C_1, C_2, C_3, C_4, C_5\}$ = {packet loss rate, codec,bit rate, frame rate,resolution}, and $i, j \in \{1, 2, 3, 4, 5\}$.

Table 1. Description of the 1-9 comparison scale.

Scale a_{ij}	Description
1	C_i and C_j have equal impact
3	C_i has a slightly greater impact than C_j
5	C_i has a stronger impact than C_j
7	C_i has a significantly stronger impact than C_j
9	C_i has an absolutely stronger impact than C_j
2,4,6,8	Intermediate values representing relative impact between the above levels
1, 1/2, ..., 1/9	Reciprocal values representing the relative impact of C_j to C_i

Through multiple comparisons, the pairwise comparison matrix for the criteria layer relative to the target layer is obtained as follows:

$$A = \begin{bmatrix} 1 & 5 & 6 & 8 & 9 \\ \frac{1}{5} & 1 & 4 & 6 & 8 \\ \frac{1}{6} & \frac{1}{4} & 1 & 3 & 3 \\ \frac{1}{8} & \frac{1}{6} & \frac{1}{3} & 1 & 2 \\ \frac{1}{9} & \frac{1}{8} & \frac{1}{3} & \frac{1}{2} & 1 \end{bmatrix} \tag{1}$$

The largest eigenvalue of matrix A and the corresponding eigenvector are calculated, yielding a maximum eigenvalue of $\lambda = 5.3443$.

Next, we check the consistency of matrix A. The Consistency Index (CI) is calculated using the following formula:

$$CI = \frac{\lambda - n}{n - 1} \tag{2}$$

With $n = 5$ and $\lambda = 5.3443$, we calculate $CI = 0.0861$. Referring to the numerical table of random consistency indicators, the Random Index (RI) for matrix A is determined to be $RI = 1.12$. Using the consistency index and the random consistency index, the Consistency Ratio (CR) is computed as follows:

$$CR = \frac{CI}{RI} = 0.077 \tag{3}$$

Since $CR < 0.1$, the degree of inconsistency of matrix A is within the acceptable range. Consequently, the feature vector $w = (0.5453, 0.2585, 0.1041, 0.0546, 0.0375)^T$ can be considered as a measure of the influence of each factor.

Table 2. Random consistency index numerical table.

n	**1**	**2**	**3**	**4**	**5**	**6**
RI	0	0	0.58	0.90	1.12	1.24
n	**7**	**8**	**9**	**10**	**11**	
RI	1.32	1.41	1.45	1.49	1.51	

3.2.2 3D CNN model construction

In this section, building on the influence factor weights obtained previously, we develop a two-stream 3D convolutional neural network (CNN) distortion model. The specific implementation details are outlined below:

A: Extending the 2D Convolutional Network to 3D. The transition to three dimensions is achieved by extending all filters and pooling kernels, effectively adding a temporal dimension to the existing 2D

network. This modification converts the filter from $N \times N$ to $N \times N \times N$. Given the linear relationship between video sequences, the filter weights in 2D can be repeated N times along the time dimension and then divided by N, ensuring consistent dimensionality for the convolution filter. Since the output of the video convolutional layer, composed of frames, remains constant in the time dimension, the average and maximum pooling layers align with the 2D network structure [20].

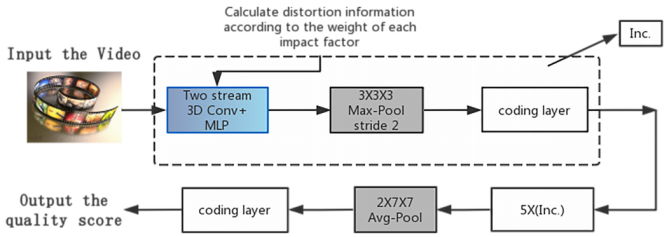


Figure 2. Network structure of the Multi-Parameter Quality Assessment (MQA) model.

B: Considering Growth Rates of Space, Time, and Network Depth. For all frames in the video stream, the spatial structure (horizontal and vertical) remains unchanged, so their pooling kernel and step size should also be consistent. This means that deeper spatial information is equally influenced by frames that are progressively farther away. However, due to temporal factors, using the same conditions does not always yield the expected results because of the impact of frame rate and image dimension. If the temporal domain grows too quickly compared to the spatial domain, capturing fine-grained distortion information becomes challenging. Conversely, if temporal growth is too slow, the degree of video distortion may be inaccurately assessed.

C: Two-Stream 3D CNN Structure. Although a 3D convolutional neural network can directly learn distortion information from continuous video frames, it still performs pure feedforward calculations. The optical flow algorithm, which is somewhat periodic (e.g., it can iteratively optimize the optical flow field), is also integrated. Therefore, a two-stream structure is adopted. One stream processes RGB stream distortion information, while the other processes optimized smooth optical flow distortion information, averaging the two results for improved accuracy.

D: Calculation of Spatial and Temporal Distortion Information. Both spatial and temporal distortion information are calculated using the 3D CNN and a feedforward neural network (MLP). This calculation determines the deviation between the network's actual output and the expected output, resulting in a quality score for the video [24].

A Multilayer Perceptron (MLP) is a type of feedforward neural network that includes an input layer, at least one hidden layer, and an output layer. In practical applications, an activation function is typically chosen to enhance adaptability for classification tasks. MLPs are widely used for prediction, classification, and recognition tasks. They can solve non-linear separable problems, perform iterative learning, load datasets into the network sequentially, and adjust the weights associated with the input values each time.

We use $w = (0.5453, 0.2585, 0.1041, 0.0546, 0.0375)^T$ as the weights for extracting distortion information in the two-stream 3D CNN. The delta rule is applied to update the input weight, minimizing the error in the neural network output through gradient descent. The error value for a single output neuron is a function of its actual value and target value. The total error of the network is the sum of all error values from all output neurons.

$$\text{Error}_{\text{total}} = \sum_{i=1}^{n} \frac{1}{2} (\text{tar}_i - \text{act}_i)^2 \tag{4}$$

E: Training and Activation Function. During the training phase, we utilize the rectified linear unit (ReLU) as the activation function and evaluate its performance based on the measured values of SRCC and PLCC. For detailed information about SRCC and PLCC, please refer to Section 4.2. Each activation function is gradually implemented into the hidden layer, with 100 nodes per hidden layer. Accuracy is calculated based on the closeness of the actual model output to the target output. The purpose of employing the hidden layer is to better abstract the degree of distortion for each influencing factor.

F: Encoding Layer for Error Extraction. An encoding layer is defined to extract meaningful errors from each layer of the deep CNN model, eliminating redundant information. These errors are then connected to form a comprehensive distortion score.

G: Preventing Overfitting. To prevent overfitting and improve model generalization, we employ dropout and k-fold cross-validation methods. The learning rate is set to $I = 0.00163$. The model undergoes training 100 times, with each session using 800 videos.

H: Loss Function Construction. The loss function is formulated to train the model:

$$\text{Loss} = \frac{\sum_{i=1}^{n} \|f(x_i, w) - y_i\|}{n} \tag{5}$$

where n represents the number of videos, w denotes a parameter in the network, y_i is the tagged score of the video, and $f(x_i, w)$ is the predicted score of the network.

I: Weight Calculation in Convolutional Layers. In each convolutional layer, weight calculations are performed for factors influencing video quality, including data packet loss, frame rate, bit rate, and resolution, ensuring a comprehensive evaluation of video quality.

4 Experiments

The simulation experiment environment for this study is configured as follows:

The processor used is an Intel(R) Core(TM) i5-8265U CPU @ 1.6GHz, paired with an NVIDIA GeForce MX230 graphics card. The system is equipped with 8GB of memory and runs on the Ubuntu 16.04 operating system. The development platforms utilized are TensorFlow 1.14.0 and MATLAB R2016a. Due to hardware constraints, the server configuration mirrors that of the client configuration.

4.1 Datasets

We utilized the LIVE [6] and CSIQ datasets to validate the performance of our model. The LIVE dataset comprises 160 videos, including 10 lossless videos, each associated with 15 videos exhibiting various types and levels of distortion. These distortions include MPEG-2 compression, H.264 compression, IP network distortion, and wireless network distortion. Each video is assigned a subjective score by dozens of subjects, with scores ranging from 0 to 100, where higher scores indicate poorer video quality.

The CSIQ Subjective Video Quality Database consists of 228 videos, including 12 original videos and 216 distorted ones. All videos are in YUV420 format, with varying frame rates of 24, 25, 30, 50, and 60 FPS, and have a duration of 10 seconds. The distortion types in this dataset include four compression-based distortions—H.264 compression (H.264), HEVC/H.265 compression (HEVC), Dynamic JPEG compression (MJPEG), wavelet-based compression using the SNOW codec (SNOW)—and two transmission-based distortions: H.264 video with analog wireless transmission loss (Wireless), and Additive White Gaussian Noise (AWGN).

To mitigate the risk of overfitting, the datasets were divided into training, testing, and cross-validation sets in specific ratios. Using the LIVE and CSIQ datasets, we generated a total of 32,079 distortion samples for training, validating, and testing the neural networks. The validation set was primarily used to assess the neural network's ability to predict perceived video quality accurately. The training and validation processes were repeated 10 times, utilizing 10-fold cross-validation to ensure the robustness and reliability of the average accuracy score.

By incorporating such extensive validation measures, we aim to establish the credibility of our model and demonstrate its effectiveness in various scenarios. The inclusion of multiple types and levels of distortions in the datasets allows us to rigorously test the model's performance and its ability to generalize across different video quality issues. This comprehensive approach ensures that our model is not only accurate but also resilient to a wide range of real-world conditions.

4.2 Evaluation Index

To assess the effectiveness of the proposed video quality algorithm, we utilize Spearman's Rank Order Correlation Coefficient (SRCC) and Pearson's Linear Correlation Coefficient (PLCC). These metrics evaluate the correlation between a set of estimated visual quality scores and the human subjective quality scores, as described below:

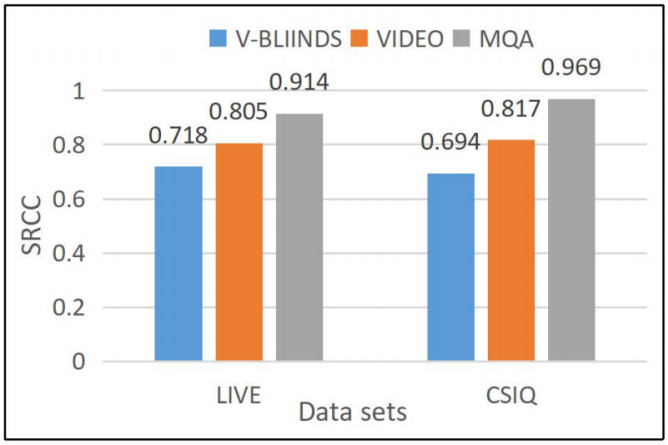
$$\begin{aligned} \text{SRCC}(Q_{\text{est}}, Q_{\text{sub}}) &= 1 - \frac{6\sum d_i^2}{m(m^2-1)} \\ \text{PLCC}(Q_{\text{est}}, Q_{\text{sub}}) &= \frac{\text{cov}(Q_{\text{est}}, Q_{\text{sub}})}{\sigma(Q_{\text{est}})\sigma(Q_{\text{sub}})} \end{aligned} \tag{6}$$

where m is the number of videos in the database, and d_i is the rank difference of the evaluation sample with serial number i in the two evaluation scores. For both metrics, values closer to 1 indicate better measurement performance. PLCC measures the degree of linear correlation between videos, while SRCC evaluates the predictive monotonicity of the algorithm.

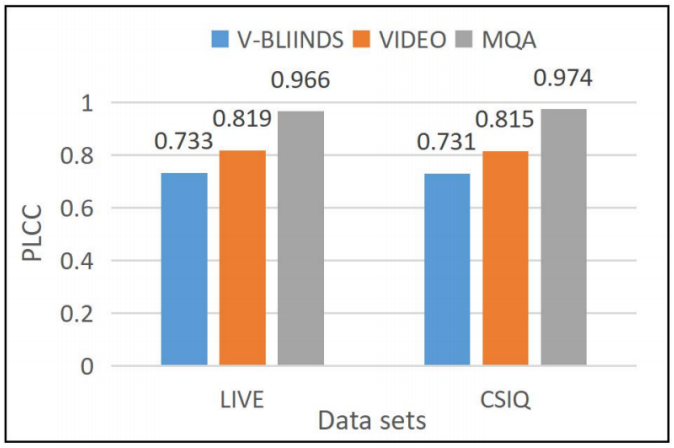
4.3 Performance Test

We first used the MQA scheme to evaluate each category of sub-distortion in the two datasets and

then conducted an overall video quality assessment, comparing the performance with the V-BLIINDS and VIDEO quality assessment algorithms. To ensure the robustness of the results, each calculation was repeated 10 times. The results are shown in the tables below.



(a)



(b)

Figure 3. Performance comparison of various schemes under SRCC and PLCC evaluation indicators.

An analysis of the tables above indicates that when using SRCC and PLCC coefficients to quantify video quality, values closer to 1 represent better visual fidelity to human perception. It is apparent that the SRCC and PLCC coefficients of the proposed scheme are generally higher than those of the V-BLIINDS and VIDEO schemes, with most values surpassing 0.9. However, in the LIVE dataset, the MQA scheme shows a slightly lower SRCC coefficient for wireless network distortion compared to the V-BLIINDS scheme. This discrepancy may be due to the calculated weights not accurately representing the type of wireless network distortion. Figure 3 illustrates the overall video quality evaluation comparison among the MQA, V-BLIINDS, and VIDEO schemes, where the MQA scheme's index values consistently exceed 0.9. In summary, whether evaluated from the perspective of prediction accuracy or monotonicity, and whether considering individual distortion subsets or the overall dataset, the proposed MQA scheme demonstrates superior performance.

Additionally, we assigned varying packet loss rates to the 12 original videos in the CSIQ dataset and constructed a corresponding set of distorted videos. This set comprises 12 groups, each containing 6 distorted videos corresponding to 6 different packet loss rate settings. We calculated the SRCC and PLCC for each packet loss rate within the 12 groups and averaged the results. To ensure robustness, the calculations were repeated 10 times, and the average value was used to assess changes in video quality under different packet loss rates.

As shown in Figure 4, with a gradual increase in packet loss rate, the SRCC and PLCC values for all three schemes exhibit a downward trend. The changes in the MQA scheme are more stable and closer to 1 compared to the other schemes. In contrast, the SRCC and PLCC indicators for the VIDEO and V-BLIINDS schemes initially decline slowly and then sharply as the packet loss rate increases. This trend indicates that under conditions of packet loss distortion, the MQA scheme provides a more reliable video quality assessment than the other two schemes.

Furthermore, our analysis highlights the resilience and robustness of the MQA scheme in handling different types of distortions and varying packet loss rates. The superior performance of the MQA scheme can be attributed to its sophisticated approach in capturing both spatial and temporal distortion features, as well as its effective use of advanced neural network techniques. By consistently delivering high accuracy in video quality assessments, the MQA scheme proves to be a valuable tool for optimizing video quality management in cloud storage systems. These findings underscore the potential of the MQA scheme to enhance user experience by ensuring high-quality video delivery even in challenging network conditions.

To evaluate the model's performance, we utilized videos from the LIVE dataset. We selected 10 lossless videos to serve as reference videos stored on the cloud server and used all corresponding distorted videos as client videos to create 10 test groups. The quality score for each group, including both lossless and distorted videos, was calculated using SRCC and PLCC as evaluation metrics. The SRCC and PLCC values for all distorted videos were averaged to derive the final

Table 3. Quality assessment results for the LIVE dataset.

Indicators	Schemes	MPEG-2	H.264	IP	Wireless	Total
SRCC	V-BLIINDS	0.620	0.749	0.692	0.873	0.718
	VIDEO	0.801	0.775	0.711	0.814	0.805
	MQA	0.972	0.930	0.994	0.869	0.914
PLCC	V-BLIINDS	0.781	0.709	0.644	0.851	0.733
	VIDEO	0.802	0.722	0.681	0.899	0.819
	MQA	0.953	0.935	0.930	0.972	0.966

Table 4. Quality assessment results for the CSIQ dataset.

Indicators	Schemes	H.264	HEVC	MJPEG	SNOW
SRCC	V-BLIINDS	0.625	0.533	0.650	0.712
	VIDEO	0.758	0.710	0.629	0.883
	MQA	0.953	0.970	0.939	0.963

Indicators	Schemes	Wireless	AWGN	Total
SRCC	V-BLIINDS	0.648	0.715	0.694
	VIDEO	0.873	0.749	0.817
	MQA	0.971	0.943	0.969

(a)

Indicators	Schemes	H.264	HEVC	MJPEG	SNOW
PLCC	V-BLIINDS	0.649	0.746	0.592	0.798
	VIDEO	0.735	0.830	0.591	0.622
	MQA	0.956	0.972	0.959	0.982

Indicators	Schemes	Wireless	AWGN	Total
PLCC	V-BLIINDS	0.705	0.897	0.731
	VIDEO	0.873	0.749	0.815
	MQA	0.977	0.926	0.974

(b)

client SRCC and PLCC values, as shown in Table 5. A quality assessment model is considered effective if the server-side video's SRCC and PLCC parameters are higher. As demonstrated in the table below, all test groups meet these criteria.

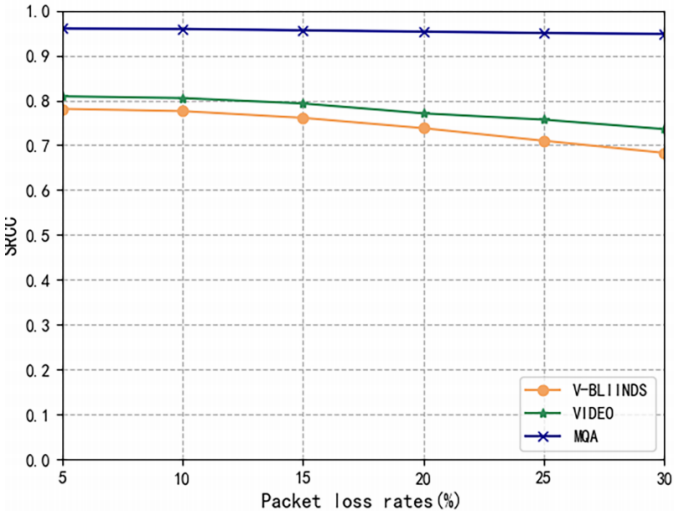
By selecting a diverse set of 10 lossless reference videos and their corresponding distorted versions, we ensured a comprehensive evaluation of the model's ability to handle various types of video distortions. This approach allowed us to rigorously test the model's performance across different scenarios and distortion levels. The use of SRCC and PLCC as evaluation metrics provided a robust framework for quantifying the correlation between the predicted video quality scores and the subjective human assessments, ensuring that the model's predictions align with human perception.

The results, summarized in Table 5, indicate that our model consistently achieves high SRCC and PLCC values for the server-side videos, demonstrating its effectiveness in video quality assessment. The consistency of these results across all test groups highlights the model's robustness and reliability in different contexts. This thorough evaluation underscores the model's capability to accurately assess video quality, making it a valuable tool for optimizing video management in cloud storage systems.

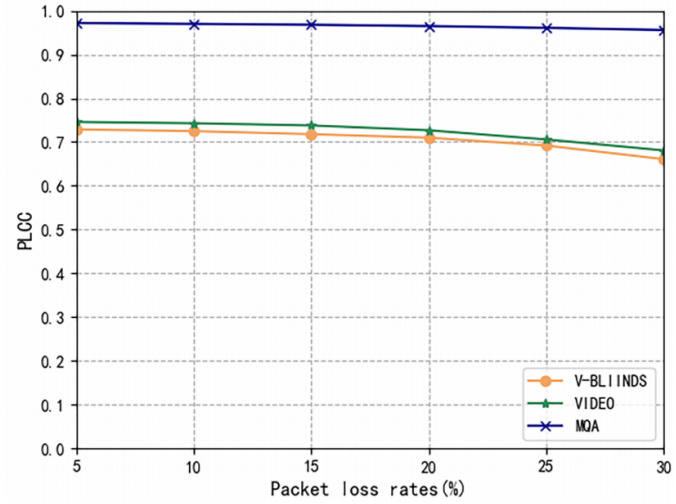
Moreover, the detailed analysis of the results provides insights into the strengths and potential areas for improvement of the model. By examining the specific cases where the model performs exceptionally well or encounters challenges, we can refine our approach and enhance the model's performance further. This iterative process of evaluation and refinement is crucial for developing a state-of-the-art video quality assessment model that meets the demands of modern cloud-based video storage and streaming services.

Table 5. Client and server video quality assessment.

Groups	Server		Client		If $Q_S > Q_C$
	SRCC	PLCC	SRCC	PLCC	
1	0.953	0.977	0.949	0.958	Yes
2	0.916	0.928	0.903	0.925	Yes
3	0.947	0.951	0.943	0.949	Yes
4	0.972	0.988	0.961	0.960	Yes
5	0.949	0.967	0.938	0.954	Yes
6	0.991	0.989	0.985	0.987	Yes
7	0.962	0.984	0.953	0.970	Yes
8	0.979	0.951	0.967	0.949	Yes
9	0.984	0.935	0.971	0.903	Yes
10	0.993	0.976	0.990	0.962	Yes



(a)



(b)

Figure 4. Performance comparison under different packet loss rates using SRCC and PLCC evaluation indicators.

4.4 Time cost

We also utilized the videos from the LIVE dataset in the aforementioned performance experiment to assess the model's time overhead. The server-side video and the video to be uploaded by the client remained unchanged throughout the experiment. We separately computed the calculation time for the server and client across the 10 groups, and the average calculation time for all distorted videos on the client side was used as the final result, excluding the model training time. We then compared the total time cost with that of the V-BLIINDS and VIDEO schemes. The results are presented in the table below.

The proposed MQA scheme demonstrates a significantly smaller time overhead compared to the other two schemes. Specifically, the MQA scheme can reduce the average time cost by 81.79% compared to the V-BLIINDS scheme and by 22.6% compared to the VIDEO scheme. These substantial reductions in time overhead highlight the efficiency of the MQA scheme in processing video quality assessments.

The time measurements in this study were conducted using a CPU, and it is anticipated that employing a GPU would result in even lower time overheads. The use of a GPU could further enhance the computational efficiency of the MQA scheme, making it even more suitable for large-scale video quality assessment tasks in real-time applications. By leveraging the parallel processing capabilities of GPUs, the MQA scheme could process video data more swiftly, thereby reducing latency and improving the overall user experience.

Additionally, the reduction in time overhead has significant implications for the scalability and practicality of the MQA scheme. In a cloud-based environment where numerous videos are uploaded and processed simultaneously, minimizing the time required for quality assessment is crucial. The MQA scheme's ability to deliver fast and accurate video quality assessments makes it an ideal choice for cloud service providers looking to optimize their video management systems.

In summary, the MQA scheme not only offers superior accuracy in video quality assessment but also ensures that the process is completed in a timely manner. The results from the time overhead analysis underscore the scheme's potential to enhance the efficiency of video quality assessment processes, thereby contributing to better resource utilization and improved service delivery in cloud environments.

Table 6. Time cost of MQA scheme.

Groups	Time cost of Server (s)	Time cost of Client (s)	Total (s)
1	49	66	115
2	65	68	133
3	42	46	88
4	45	49	94
5	60	66	126
6	51	58	109
7	69	74	143
8	47	50	97
9	51	54	105
10	44	49	93

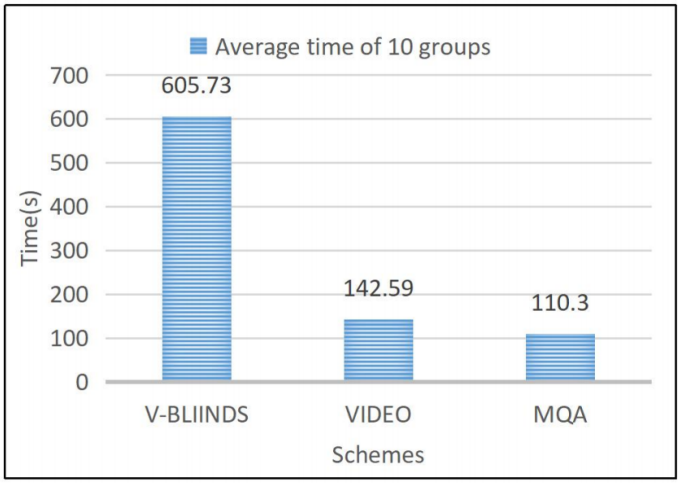


Figure 5. Comparison of the time cost of the three schemes.

5 Conclusions and Discussions

In this paper, we present an effective video quality assessment model and establish an interactive framework for quality assessment between the client and the server. First, we employ a method similar to the Analytic Hierarchy Process (AHP) to estimate the impact of packet loss rate, codec, frame rate, bit rate, and resolution on video quality by calculating the weight vector. Next, we capture the details of video distortion using a 3D convolutional neural network, focusing on both spatial and temporal flows. The coding layer is defined to remove redundant distortion information, while dropout and k-fold cross-validation methods are utilized to prevent overfitting. We then construct a loss function to train the network and output the video quality score. For model assessment, we use Spearman's Rank Order Correlation Coefficient (SRCC) and Pearson's Linear Correlation Coefficient (PLCC) and validate the model using the LIVE and CSIQ datasets. Experimental results indicate that the proposed scheme is more effective than the V-BLIINDS and VIDEO schemes in evaluating video quality. Additionally, we test the performance of the three schemes under various packet loss rate settings. A portion of the dataset is used to simulate the interactive quality assessment process between the client and the server, with test results aligning with the expected outcomes. Finally, we evaluate the time overhead of the MQA scheme, excluding the model training time, demonstrating that it is more time-efficient than the V-BLIINDS and VIDEO schemes.

The proposed video quality assessment scheme effectively evaluates the quality of videos from both the client and cloud server. However, there are still two potential research directions: Firstly, for non-uniformly distorted videos, the degree of distortion varies across small video frame blocks, necessitating a more fine-grained approach to capture video distortion information using fixed-size blocks. Secondly, constructing supervision information can enhance the network's learning ability in scenarios with limited labeled data, thereby increasing the network's adaptability. In future work, we will focus on these two research aspects.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgement

This work was supported without any funding.

References

[1] Yang, X., Lu, R., Choo, K. K. R., Yin, F., & Tang, X. (2017). Achieving efficient and privacy-preserving cross-domain big data deduplication in cloud. *IEEE Transactions on Big Data*, 8(1), 73-84. [CrossRef]

[2] Wu, X., Hauptmann, A. G., & Ngo, C. W. (2007, September). Practical elimination of near-duplicates from web video search. In *Proceedings of the 15th ACM international conference on Multimedia* (pp. 218-227). [CrossRef]

[3] Yao, J. Y., & Liu, G. (2018). Bitrate-based no-reference video quality assessment combining the visual perception of video contents. *IEEE Transactions on Broadcasting*, 65(3), 546-557. [CrossRef]

[4] Botia Valderrama, D. J. L., & Gaviria Gómez, N. (2016). Nonintrusive method based on neural networks for video quality of experience assessment. *Advances in Multimedia*, 2016(1), 1730814. [CrossRef]

[5] Søgaard, J., Forchhammer, S., & Korhonen, J. (2015, May). Video quality assessment and machine learning: Performance and interpretability. In *2015 Seventh International Workshop on Quality of Multimedia Experience* (QoMEX) (pp. 1-6). IEEE. [CrossRef]

[6] Seshadrinathan, K., Soundararajan, R., Bovik, A. C., & Cormack, L. K. (2010, February). A subjective study to evaluate video quality assessment algorithms. In *Human Vision and Electronic Imaging XV* (Vol. 7527, pp. 128-137). SPIE. [CrossRef]

[7] Saad, M. A., Bovik, A. C., & Charrier, C. (2014). Blind prediction of natural video quality. *IEEE Transactions on image Processing*, 23(3), 1352-1365. [CrossRef]

[8] Mittal, A., Soundararajan, R., & Bovik, A. C. (2012). Making a "completely blind" image quality analyzer. *IEEE Signal processing letters*, 20(3), 209-212. [CrossRef]

[9] Loh, W. T., & Bong, D. B. L. (2018). A just noticeable difference-based video quality assessment method with low computational complexity. *Sensing and Imaging*, 19, 1-20. [CrossRef]

[10] Cheng, Z., Ding, L., Huang, W., Yang, F., & Qian, L. (2017, June). A unified QoE prediction framework for HEVC encoded video streaming over wireless networks. In *2017 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting (BMSB)* (pp. 1-6). IEEE. [CrossRef]

[11] Anegekuh, L., Sun, L., Jammeh, E., Mkwawa, I. H., & Ifeachor, E. (2015). Content-based video quality prediction for HEVC encoded videos streamed over packet networks. *IEEE Transactions on Multimedia*, 17(8), 1323-1334. [CrossRef]

[12] Alreshoodi, M., Adeyemi-Ejeye, A. O., Woods, J., & Walker, S. D. (2015). Fuzzy logic inference system-based hybrid quality prediction model for wireless 4kUHD H. 265-coded video streaming. *IET Networks*, 4(6), 296-303. [CrossRef]

[13] Zhang, Y., Gao, X., He, L., Lu, W., & He, R. (2018). Blind video quality assessment with weakly supervised learning and resampling strategy. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(8), 2244-2255. [CrossRef]

[14] Valderrama, J. F. B., & Valderrama, D. J. L. B. (2018). On LAMDA clustering method based on typicality degree and intuitionistic fuzzy sets. *Expert Systems with Applications*, 107, 196-221. [CrossRef]

[15] Li, Y., Po, L. M., Cheung, C. H., Xu, X., Feng, L., Yuan, F., & Cheung, K. W. (2015). No-reference video quality assessment with 3D shearlet transform and convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(6), 1044-1057. [CrossRef]

[16] Agarla, M., Celona, L., & Schettini, R. (2020). No-reference quality assessment of in-capture distorted videos. *Journal of Imaging*, 6(8), 74. [CrossRef]

[17] Nightingale, J., Salva-Garcia, P., Calero, J. M. A., & Wang, Q. (2018). 5G-QoE: QoE modelling for ultra-HD video streaming in 5G networks. *IEEE Transactions on Broadcasting*, 64(2), 621-634. [CrossRef]

[18] Narwaria, M., & Lin, W. (2011, September). Machine learning based modeling of spatial and temporal factors for video quality assessment. In *2011 18th IEEE International Conference on Image Processing* (pp. 2513-2516). IEEE. [CrossRef]

[19] Pal, D., & Vanijja, V. (2017). A No-Reference Modular Video Quality Prediction Model for H. 265/HEVC and VP9 Codecs on a Mobile Device. *Advances in Multimedia*, 2017(1), 8317590. [CrossRef]

[20] Qian, L., Pan, T., Zheng, Y., Zhang, J., Li, M., Yu, B., & Wang, B. (2020). No-Reference Nonuniform Distorted Video Quality Assessment Based on Deep Multiple Instance Learning. *IEEE MultiMedia*, 28(1), 28-37. [CrossRef]

[21] Chen, P., Li, L., Ma, L., Wu, J., & Shi, G. (2020, October). RIRNet: Recurrent-in-recurrent network for video quality assessment. In *Proceedings of the 28th ACM international conference on multimedia* (pp. 834-842). [CrossRef]

[22] Li, D., Jiang, T., & Jiang, M. (2021). Unified quality assessment of in-the-wild videos with mixed datasets training. *International Journal of Computer Vision*, 129(4), 1238-1257. [CrossRef]

[23] Zhen, P., Chen, H. B., Cheng, Y., Ji, Z., Liu, B., & Yu, H. (2021). Fast video facial expression recognition by a deeply tensor-compressed LSTM neural network for mobile devices. *ACM Transactions on Internet of Things*, 2(4), 1-26. [CrossRef]

[24] Wang, N., Fang, F., & Feng, M. (2014, May). Multi-objective optimal analysis of comfort and energy management for intelligent buildings. In *The 26th Chinese control and decision conference (2014 CCDC)* (pp. 2783-2788). IEEE.

[25] Fang, F. A. N. G., Tan, W., & Liu, J. Z. (2005). Tuning of coordinated controllers for boiler-turbine units. *Acta Automatica Sinica*, 31(2), 291-296.

[26] Fang, F., Jizhen, L., & Wen, T. (2004). Nonlinear internal model control for the boiler-turbine coordinate systems of power unit. *PROCEEDINGS-CHINESE SOCIETY OF ELECTRICAL ENGINEERING*, 24(4), 195-199.

[27] Lv, Y., Fang, F. A. N. G., Yang, T., & Romero, C. E. (2020). An early fault detection method for induced draft fans based on MSET with informative memory matrix selection. *ISA transactions*, 102, 325-334. [CrossRef]