Blind Video Quality Assessment based on Spatio-Temporal Feature Resolver

Xiaodong Bi[a], Xiaohai He[b,*], Shuhua Xiong[b], Zeming Zhao[b], Honggang Chen[b], Raymond Edward Sheriff[c]

[a]School of Cyber Science and Engineering, Sichuan University, Chengdu 610065, China
[b]College of Electronics and Information Engineering, Sichuan University, Chengdu 610065, China
[c]Department of Computer Science, Edge Hill University, Ormskirk, Lancashire, L39 4QP, United Kingdom

Abstract

As a method to measure video quality, blind video quality assessment (BVQA) plays an important role in video related applications. Feature extraction, as a vital component of BVQA, significantly impacts the performance and speed of the method. In the process of designing BVQA method, we make the feature extraction process independent of the whole quality evaluation method. First, a Spatio-Temporal Feature Resolver (STFR) is obtained by training. Then, the STFR is employed to directly extract the spatio-temporal features of the video sequences. Finally, the extracted features are mapped to quality scores using Support Vector Regression (SVR). STFR only needs to be trained in an undistorted video sequence, and can be directly applied to video sequences of various scenes, which has a universality. To evaluate the effectiveness of the proposed method, the experimental results on four published video quality databases show that the proposed BVQA method not only achieves more accurately than the existing BVQA methods, but also exhibits significant competitiveness in computational speed.

Keywords: Blind Video Quality Assessment, Spatio-Temporal Feature Resolver, Support Vector Regression, Universality

1. Introduction

With the rapid development of internet technology and the widespread utilization of multimedia equipment, digital multimedia information has become increasingly prevalent in our daily life. Video, as an important branch of digital media, contains more appropriate and comprehensive information compared to other forms of digital media [1, 2]. However, the video information undergoes lossy compression during processing, and these videos are also susceptible to various distortions during storage, transmission, and reproduction. Any of these distortions can adversely affect the video quality. Therefore, video quality assessment (VQA) is crucial in related fields of video processing [3]. The existing VQA methods are mainly divided into subjective VQA and objective VQA. Subjective VQA methods involve directly scoring video content based on human subjective perception. Although subjective VQA is accurate and reliable, it is time-consuming and laborious. Therefore, it is very necessary to design an objective VQA method that aligns with human visual perception characteristics for the study of video processing [4, 5].

Based on the extent to which reference information is utilized in the VQA methods, objective VQA can be divided into three types: Full-Reference VQA (FR-VQA) [3, 6–8], Reduced-Reference VQA (RR-VQA) [9], and No-Reference VQA (NR-VQA) or Blind VQA (BVQA) [10–19]. Due to the unavailability of perfect videos in real-life scenarios and the unknown distortions introduced during video shooting, compression and other video processing stages, FR-VQA and RR-VQA methods are not applicable [13]. In contrast, BVQA holds more realistic research significance and practical value as it doesn't rely on reference video information [20].

The performance of BVQA methods largely depends on the effective extraction of video features, and the utilization of effective features is beneficial for enhancing performance [21, 22]. In the early stage of BVQA methods research, video frames were regarded as independent images, and the blind image quality assessment (BIQA) methods were used to calculate the quality scores of individual video frames. Finally, all the video frames were statistically calculated to obtain the final BVQA result. There exist both similarities and differences between BVQA and BIQA [23, 24]. Unlike BIQA, which only considers spatial information, BVQA needs to account for both spatial and temporal information, as temporal information is unique and crucial for video content. Simply applying BIQA methods to BVQA without properly extracting temporal features may adversely affect the quality assessment results of videos [25].

To further improve the performance of BVQA methods, there is a integration of spatial features extracted from BIQA methods with manually extracted temporal features from the video sequence [26–28]. Despite the exploration and design of manually extracted spatio-temporal features in various scenarios, the diversity and complexity of video information content, the generalization ability of these BVQA methods is not strong and there are limitations [29].

In recent years, deep learning-based BVQA methods have been gained popularity and have significantly improved performance compared to traditional BVQA methods. However, these deep learning-based BVQA methods require a large-scale

*Corresponding author:Xiaohai He
Email address: hxh@scu.edu.cn (Xiaohai He)

labeled dataset and considerable training time [30]. While some of these methods utilize weak labeling from FR-VQA or RR-VQA to increase the large-scale labeled data [29–31], others speed up the computation by designing lightweight networks [19, 32]. Moreover, these methods may suffer from bias and inadequacy. This can be a challenge, especially when BVQA is employed as a performance assessment method in other video processing applications, as the computational complexity and time-consuming nature can impact widespread adoption [33]. Therefore, there is an urgent need to design simple and efficient BVQA methods.

In order to address the limitations of manual feature extraction methods and the computing time-consuming of deep learning methods in feature extraction process. We are committed to effectively extracting spatio-temporal features from video sequences of various scenes. In pursuit of this goal, we train a general model to extract video spatio-temporal features by learning the statistical properties of video information. To effectively extract spatio-temporal features from video sequences of various scenes and accurately calculate video quality scores, we employ a two-step method. Firstly, we have developed an independent Spatio-Temporal Feature Resolver (STFR) using advanced learning techniques. Secondly, the STFR is used to efficiently and automatically extract the spatio-temporal features from video sequences, and we employ a regression model called Support Vector Regression (SVR) to establish the mapping between the extracted visual features and quality scores. This BVQA method is named as STFR-BVQA, highlighting the significance of the STFR in the overall BVQA process.

It is worth mentioning that the STFR can be directly applied to the spatio-temporal feature extraction of cross-scene video through the training of video sequences, which has a universality. We conducted a series of experiments on four publicly available databases. The results clearly demonstrate that STFR-BVQA outperforms existing BVQA methods in terms of both performance and computational speed. The main contributions of this work are as follows:

- Aiming at BVQA, we have proposed a visual model called STFR that enables the automatic extraction of video spatio-temporal features. As a model, STFR can not only automatically extract video spatio-temporal features, but also automatically apply to videos of different scenes after only one training.

- The STFR is not only universal, but also can be obtained with only one video sequence. Meanwhile, the time of extracting video spatio-temporal features can be saved greatly by using STFR.

- We propose a BVQA method based on STFR, which takes STFR as an independent model to extract video spatio-temporal features. This method takes into account the impact of both similarity and difference in video features on video perceived quality.

- Through extensive experiments conducted on four wild video databases, the experimental results clearly demonstrate that the STFR-BVQA method has significant advantages over existing methods in terms of quality assessment accuracy and computational processing efficiency.

The paper is structured as follows: Section II introduces the related work of BIQA and BVQA. Section III presents the process of the proposed BVQA method on the basis of STFR. Section IV describes the experimental results and analysis. Finally, a conclusion is drawn in Section V.

2. Relate Work

In this section, we will present the related work on BIQA and BVQA. Existing BVQA methods have been influenced by BIQA techniques. Therefore, we will first introduce the work on BIQA and then proceed to discuss the work on BVQA.

2.1. Blind Image Quality Assessment

According to the way of image feature extraction, existing BIQA methods can be categorized into manual feature extraction and deep learning based BIQA [34, 35]. The BIQA based on manual feature extraction utilizes the statistical features of images obtained through statistical observations in the spatial domain by manually extracting the image features, typically represented by specific probabilistic model parameters [36–40]. In 2011, Moorthy et al. [36] designed a BIQA method to extract image features from the discrete wavelet transform (DWT). They leveraged the transformation of spatial information into the transform domain to enable BIQA. Similarly, in 2012, Saad et al. [37] utilized the coefficients of a statistical model based on image discrete cosine transform (DCT) as image features for BIQA. In 2012, Mittal et al [38] developed a BIQA methods by extracting statistical features from nature scene images in the spatial domain. This method achieves BIQA by constructing a statistical model of natural image. In 2013, Mittal et al. [39] introduced a completely blind image quality analyzer. This method aims to evaluate image quality without relying on any reference image information and Mean Opinion Score (MOS) values. Similarly, in 2015, Zhang et al. [40] proposed a feature-enriched completely blind image quality evaluator. These two methods belong to the completely BIQA, which is carried out the quality evaluation in an unsupervised manner. While these methods based on manual feature extraction have achieved impressive performance on synthesized distorted images, their effectiveness is limited when dealing with the complexity of authentic distortion types and image content [41–46]. The diversity and variability of authentic distortions pose challenges for these methods, making it difficult to accurately evaluate the quality of real images. Therefore, there is a need for more robust and adaptive methods that can handle a wide range of distortion types and image variations.

Recently, deep learning-based BIQA methods have gained significant attention and have demonstrated superior prediction performance compared to traditional BIQA methods [47]. The utilization of deep learning techniques allows to effectively learn complex and discriminative features from images, enabling more accurate and robust BIQA [48–53]. In 2014, Kang

et al. [48] introduced convolutional neural networks (CNN) into the BIQA method for the first time. In 2018, Ma et al. [49] designed a unified end-to-end neural network (NN) framework to jointly learn image classification and BIQA tasks. In 2020, Zhang et al. [50] designed two linear CNNs for synthetic distorted images and authentic distorted images respectively, and combined them to achieve BIQA. In 2021, Ma et al. [51] implemented BIQA metrics by using generative adversarial networks (GAN) to simulate the active reasoning process of the internal generative mechanism. In 2022, Pan et al. [52] introduced a vision compensation recovery network based on GAN, aiming to enhance image quality assessment accuracy across multiple databases. In 2023, Liu et al. [53] decomposed the image into multiple spatial frequency bands through multi-scale filtering to capture global and local details, and then map image features to quality scores using CNN.

One of the main challenges in developing deep learning-based BIQA methods is the insufficient availability of large-scale labeled datasets for training network models [54]. The existing datasets are too small to adequately train network models. To overcome this challenge, current deep learn-based BIQA methods employ transfer learning or progressive learning techniques [55], where network models are pre-trained on large-scale datasets such as ImageNet and subsequently fine-tuned for the specific task of BIQA [56, 57]. While this method has achieved certain results, it requires careful selection of certain hyperparameters, and excessive fine-tuning can lead to overfitting [34, 58].

2.2. *Blind Video Quality Assessment*

In early BVQA methods, the video sequence was divided into individual frames, the BIQA methods were applied to calculate the quality score for each video frame, and the average predicted score of all video frames was calculated as the overall quality score of the video sequence [59, 60]. However, this approach completely ignores the temporal information of the video [36, 61–64]. Based on the spatial information extracted by BIQA methods, traditional BVQA methods manually extract temporal information from videos. By integrating both spatial and temporal information to ultimately predict video quality scores [26, 65].

In 2014, Saad et al. [10] introduced a BVQA method that incorporated a spatio-temporal natural scene statistics model and a motion model to quantify motion coherency. In 2016, Mittal et al. [11] developed a completely BVQA method based on the previous BIQA methods [39] and [40]. In 2019, Korhonen et al. [12] proposed an efficient two-level video quality model. This method selects a representative subset of video frames from the low complexity features of the entire sequence and extracts the high complexity features. In 2020, Dendi et al. [13] studied a video representation that is based on a parametric statistical model by modeling the output statistics of spatiotemporal Gabor bandpass filtering. In 2020, Korhonen et al. [14] proposed a novel fusion-based model that extracts 60 features from 763 statistical feature pools used in existing BVQA methods and applied a feature selection strategy to an efficient BVQA model.

The existing BVQA methods based on manual feature extraction have limitations and struggle to achieve ideal generalization performance [28]. In contrast, an increasing number of BVQA methods based on deep learning have demonstrated superior performance. In 2016, Li et al. [15] extracted simple and effective spatio-temporal features through 3D shearlet transform, capturing the statistical characteristics of natural scenes, and realizing BVQA combined with CNN. In 2019, Li et al. [16] proposed an objective BVQA method that integrates two prominent effects (content dependence and time memory effects) of the human visual system (HVS) into a deep NNs. In 2020, Chen et al. [17] designed a BVQA framework that conforms to the quality perception of distorted video and makes quality predictions from the perspective of HVS. In 2021, Korhonen et al. [18] combined depth features extracted from a deep network model with manually extracted features to further enhance the performance of the BVQA method. In 2022, You et al. [19] demonstrated that Transformers have shown outstanding performance in analyzing temporal data and can facilitate quality assessment of video sequences. The BVQA methods based on deep learning exhibit good performance, but the feature extraction process is complex and burdensome, requiring a substantial amount of labeled samples [66].

To address the issue of limited data, data augmentation with weakly supervised learning has been introduced [30]. However, the methods in [30] and [31] adopt single BVQA methods to generate weak labels, resulting in a strong bias in the weak label itself. In order to reduce computational complexity, Vega et al. [32] presented an online BVQA method based on unsupervised deep learning. On the server side, a model based on unsupervised deep learning is used to extract features, and on the client side, no reference metric is computed by lightweight networks. You et al. [19] proposed a lightweight BVQA method by integrating locally shared attention into a global Transformer encoder. This processing method reduces the computational input, but can significantly impact the final perceived quality [29]. There are also completely BVQA methods that do not require any reference to video information [67–69]. However, the performance and application scenarios of these methods are limited.

In short, the BVQA methods based on manual feature extraction struggle to achieve optimal generalization performance [28]. The BVQA methods based on deep learning are difficult to train database-independent and BVQA specific presentation encoders due to data shortage. Moreover, existing weakly supervised BVQA methods are susceptible to bias introduced by single weak label [29, 30, 66]. Additionally, these deep learning-based methods have a complex computation process and slow computation speed, affecting their usability [31].

3. The Proposed Method

Figure.1 provides our proposed BVQA framework. The framework is mainly divided into two parts: the design of STFR and BVQA based-on STFR. The design process of STFR speeds up the speed of spatio-temporal features extraction and reduces the complexity of BVQA.

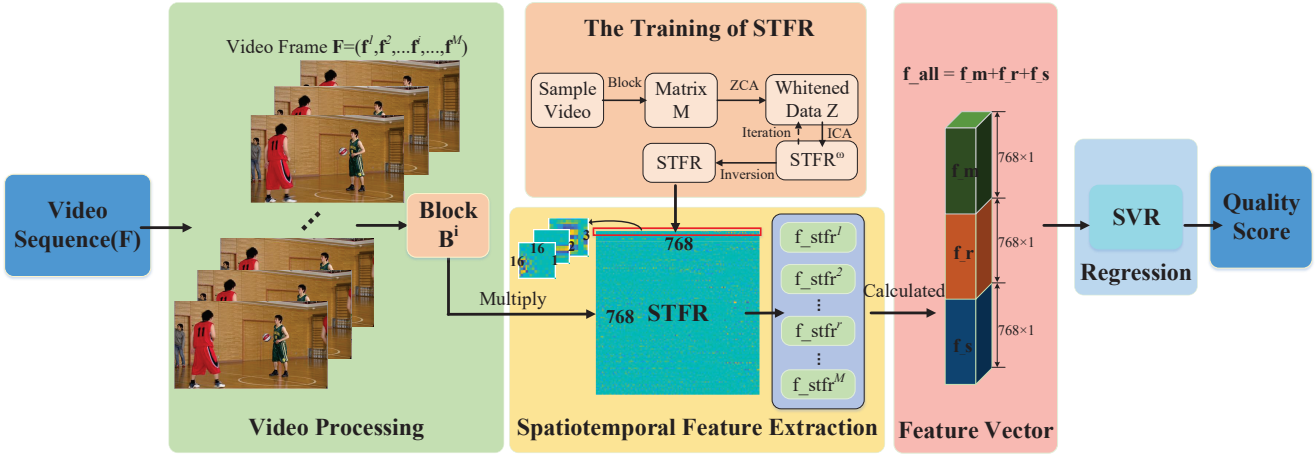

Figure 1: Framework of the proposed STFR-BVQA. The training process of STFR is independent of the whole STFR-BVQA framework.

3.1. Design of STFR

The design of STFR needs a training process. First, an undistorted video sequence is randomly selected from the video database as the sample source, and the YUV video is converted to RGB video signal, where the resolution of the video sequence is 416×240, and there are 100 video frames in total. Then, each video frame is randomly divided into 300 RGB blocks with a size of 16×16. Finally, each 16×16 RGB block is stretched into a column vector, and all the RGB blocks together form a matrix denoted as $\mathbf{M}$, with a size of 768×30000.

Next, Zero-phase Component Analysis (ZCA) is used to decompose the covariance matrix of $\mathbf{M}$ into eigenvalues and eigenvectors. The sample data is whitened to complete the preprocessing. ZCA not only reduces the redundancy of input, but also makes the whitened data map back to the source space closer to the original data. the whitening matrix is given by:

$$Z = \mathbf{V} \times \frac{1}{\mathbf{D}} \times \mathbf{V}^T \times \mathbf{M} \tag{1}$$

where $\mathbf{D}$ represents the eigenvalue of the covariance matrix, $\mathbf{V}$ represents the eigenvector, and $\mathbf{V}^T$ represents the transpose of the eigenvector.

Before the training, $STFR^{'}$ is initialized as a random matrix. Meanwhile, in order to ensure the stability and computational efficiency of the data, the STFR was orthogonalized:

$$STFR^{'} = STFR^{'} \times \frac{1}{\sqrt{STFR^{'} \times STFR^{'T}}} \tag{2}$$

Then, the Independent Component Analysis (ICA) algorithm was employed to iteratively train $STFR^{'}$ with the whitened data $\mathbf{Z}$. The iterative form of the algorithm is as follows:

$$\begin{aligned} STFR^{'}_{(n+1)} &= \mathbf{V}\{\mathbf{Z} \times g((STFR^{'}_{(n)})^T)\} \\ &-\{\mathbf{Z} \times g^{'}((STFR^{'}_{(n)})^T)\}STFR^{'}_{(n)} \end{aligned} \tag{3}$$

where $g(\cdot)$ is the hyperbolic tangent function and $g^{'}(\cdot)$ is the derivative of $g(\cdot)$.

The whole iterative process continues by repeating equations (2)-(3) until the final result converges. The convergence criteria are as follows:

$$\| \left| STFR^{'}_{(n+1)} \times STFR^{'}_{(n)}{}^T \right| - \mathbf{E} \| \leq 10^{-12} \tag{4}$$

where $\| \cdot \|$ denotes the L_2 norm, $\mathbf{E}$ stands for an identity matrix.

Finally, the obtained matrix, $STFR^{'}$ is converted from the whitening space to the original video space to obtain STFR, whose size is 768×768. Each row vector in STFR represents a video block of size 16×16×3.

3.2. Spatio-Temporal Features Extraction

In order to extract spatio-temporal features of video sequence more quickly and conveniently. The STFR is directly applied to the input video sequence. This allows for a comprehensive analysis of the spatio-temporal features, enabling the direct extraction of such features from the video sequence.

First, the input video sequence, $\mathbf{F}=(\mathbf{f}^1,\mathbf{f}^2, ..., \mathbf{f}^M)$, is preprocessed. The YUV video is converted into an RGB video signal, and the mean value of each frame is removed and divided into 16×16×3 blocks.

$$\mathbf{b}^i_j = \mathbf{f}^i_j - \mu(\mathbf{f}^i) \tag{5}$$

where $j \in (1, 2, ..., N)$, N indicates the total number of blocks of the current video frame, $i \in (1, 2, ..., M)$, M indicates the total number of video frames. $\mathbf{f}^i_j$ represents the j-th block of the i-th video frame, and $\mu(\cdot)$ represents the mean value of the current video frame. For ease of calculation and stretching each video block into a column vector, the representation of the current video frame is $\mathbf{B}^i=(\mathbf{b}^i_1,\mathbf{b}^i_2, ..., \mathbf{b}^i_N)$, the size of $\mathbf{B}^i$ is $768 \times N$.

Then, the STFR is used to extract spatio-temporal feature information from each frame of the video sequence in the form of multiplication.

$$\mathbf{f\_stfr}^{i'} = STFR \times \mathbf{B}^i \tag{6}$$

where $\mathbf{f\_stfr}^{i\prime}$ represents the visual perception features of the current video frame obtained by the STFR, and the size of $\mathbf{f\_stfr}^{i\prime}$ is $768 \times N$.

By analyzing the output signal $\mathbf{f\_stfr}^{i\prime}$, and the row standard deviation of $\mathbf{f\_stfr}^{i\prime}$ is calculated as $\mathbf{f\_stfr}^{i}$. $\mathbf{f\_stfr}^{i}$ represents a feature of video frame $\mathbf{f}^{i}$, whose size is 768×1.

Finally, the spatio-temporal feature information, **F_STFR** of the current video sequence **F** is expressed as:

$$\mathbf{F\_STFR} = (\mathbf{f\_stfr}^{1}, \mathbf{f\_stfr}^{2}, ..., \mathbf{f\_stfr}^{M}) \tag{7}$$

where **F_STFR** represents the spatio-temporal features information of video sequence **F**, and its size is $768 \times M$.

As shown in Figure.2, the line plots of two different distorted video sequences based on the video **F_STFR** obtained by the STFR are presented. As shown in Figure.2(a), it can be observed that the distribution of $\mathbf{f\_stfr}^{i}$ line plots for adjacent video frames is similar when there is no drastic change in the content between them. However, when a frame in the adjacent video frames is seriously distorted, the distribution of the distorted line plot changes dramatically while the overall distribution remains consistent, as shown in Figure.2(b). To provide a more detailed illustration, we selected five adjacent video frames and extracted 20 $\mathbf{f\_stfr}^{i}$ from each frame. Meanwhile, in order to observe the distortion changes in the curve, the third frame was synthesized by manually adding four distortion blocks. The vertical axis represents the value of $\mathbf{f\_stfr}^{i}$, while the horizontal axis represents the number of $\mathbf{f\_stfr}^{i}$.

Based on this observation, it can be concluded that the $\mathbf{f\_stfr}^{i}$ features of adjacent videos exhibit similar spatial feature information. Therefore, the row mean of **F_STFR** is calculated to represent the spatial feature of the entire video sequence.

$$\mathbf{f\_m} = \frac{1}{M}\sum_{i=1}^{n} \mathbf{f\_stfr}^{i} \tag{8}$$

where the size of M indicates the total number of video frames. **f_m** represents the mean of rows in **F_STFR**, and its size is 768×1.

Meanwhile, HVS has the temporal hysteresis effect [16]. According to the feature difference between adjacent video frames, the $\mathbf{f\_stfr}^{r}$ with the largest difference is identified as the representative frame and taken as the feature of the video sequence.

$$\mathbf{f\_r} = \mathbf{f\_stfr}^{r} = max(\mathbf{f\_stfr}^{i}, \mathbf{f\_m}) \tag{9}$$

where $\mathbf{f\_stfr}^{r}$ indicates the representative frame with the largest difference between $\mathbf{f\_stfr}^{i}$ and **f_m** in the video sequence, and is selected as the representative frame.

In the temporal feature information, when there is distortion in a certain frame of the video, there will be a big difference in $\mathbf{f\_stfr}^{r}$. Korhonoe et al. [18] also mentioned that the vector obtained by the standard deviation is the video consistency time domain eigenvector. The consistency feature effectively captures the motion consistency of the video, which is a distinctive characteristic of video information and serves as an essential component of temporal information. Therefore, the row standard deviation of **F_STFR** is calculated to serve as the representative temporal feature of the whole video sequence.

$$\mathbf{f\_s} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(\mathbf{F\_STFR}_{ij} - \mu(\mathbf{F\_STFR}_{i}))^2} \tag{10}$$

where $i \in (1, 2, ..., 768)$, i is the number of rows in **F_STFR**, $\mu(\cdot)$ represents the mean value of the i-th row in **F_STFR**. **f_s** is the standard deviation of the row in **F_STFR**, and its size is 768×1.

To better understand the relationship between extracted features and the level of distortion, we selected video sequences under four different quantization parameters (QP). The larger the QP, the higher the level of distortion in the video. Since the data values of each video feature are quite different, we chose 50 features in **f_m** and **f_s** to analyze the relationship between the features and the level of distortion respectively. As shown in Figure. 3, the horizontal axis of Figure. 3(a) represents 50 **f_m**, and the vertical axis represents the corresponding **f_m** values. The horizontal axis of Figure. 3(b) shows 50 **f_s**, and the vertical axis shows the corresponding **f_s** values. The eigenvalue constantly increases with the rise in the quantization parameter value, indicating that the eigenvalue increase with the worsening of the video distortion.

Therefore, based on **F_STFR**, **f_m**, **f_r** and **f_s** are taken as the spatio-temporal feature vector of the whole video sequence. A straightforward strategy to do this is to concatenate all features.

$$\mathbf{f\_all} = \mathbf{f\_m} \oplus \mathbf{f\_r} \oplus \mathbf{f\_s} \tag{11}$$

where $\oplus$ represents a concatenation operator, the size of **f_all** is 2304×1.

3.3. Regression

In the existing BIQA and BVQA methods, regression models such as backpropagation NN, Random Forest Regression (RFR) and SVR are commonly employed. These regression models utilize the spatio-temporal features of the video sequence as input and the corresponding MOS as the model output to train the model parameters. Korhonen et al. [18, 70] also suggest that this approach is the most feasible way to model video quality. In our experiments, we also explored different regression models (i.e., NN and RFR). However, SVR exhibited strong performance and computational efficiency [71–73]. Therefore, SVR has been as regression model in STFR-BVQA to calculate video quality score. In this study, we employ the LIBSVM package for the implement of SVR using a radial basis function kernel, and the gamma is set to 0.05.

4. Experiment Results

In this section, we first introduce the databases, experimental protocols and evaluation metrics for the experiments. Next, we compare the overall performance and cross-dataset performance of STFR-BVQA with the latest and most classical BVQA methods on the databases. Then, we also perform the

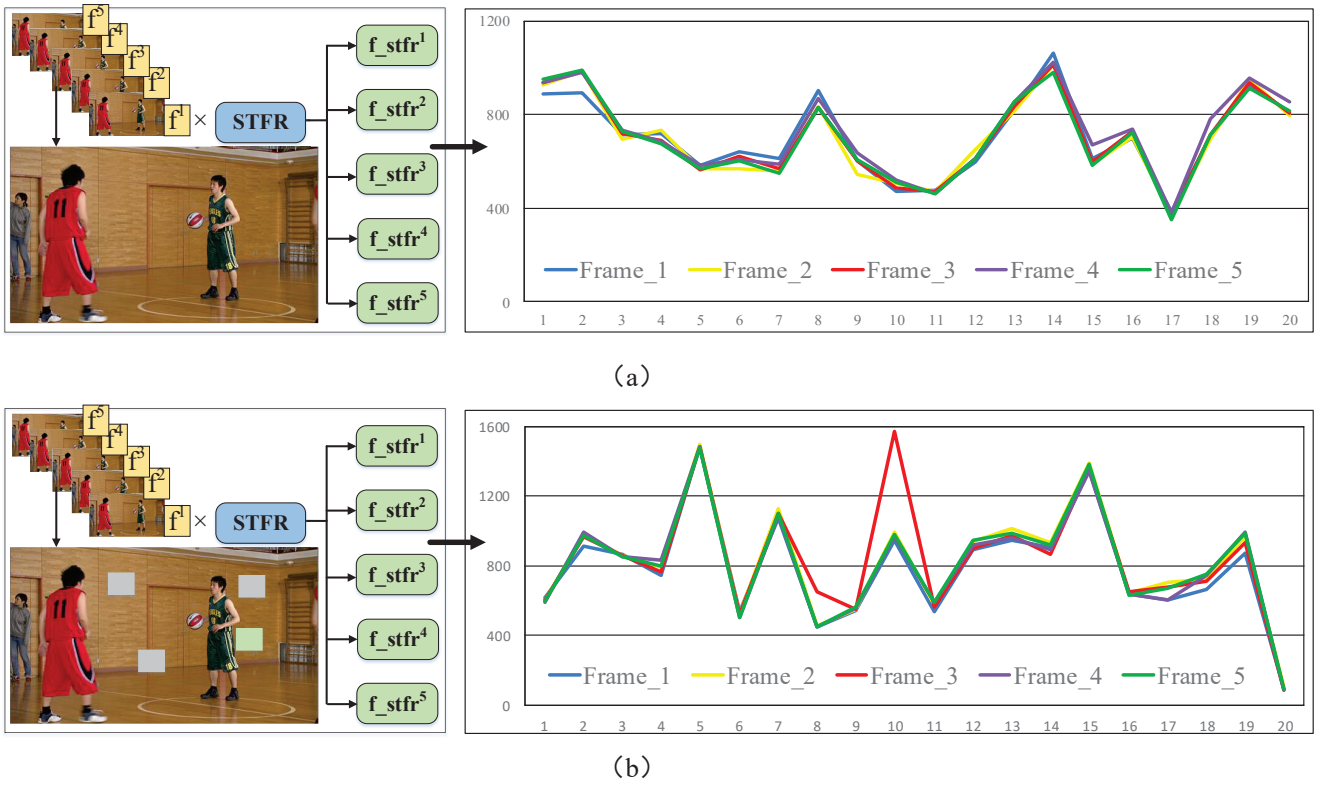

Figure 2: Curves of **F_STFR** value in video sequence. The horizontal axis represents the 20 $\mathbf{f\_stfr}^i$ in the current video frame obtained by STFR, and the vertical axis represents the corresponding STFR value. (a) **F_STFR** similarity curves of adjacent five frames. (b) **F_STFR** variation curve of five frames.

statistical significance test and analyze the computational time required by STFR-BVQA. Finally, the ablation experiment of the STFR-BVQA is carried out and STFR is replaced by some well-trained deep feature extractor. At the same time, the parameter settings in the experiment are analyzed.

4.1. Experimental Protocols and Performance Criteria

The BVQA methods should accurately reflect human perception of video quality. Therefore, the availability of a video quality database with subjective quality assessment scores is essential for constructing and validating VQA methods. In this paper, we conducted the validation experiments on four databases: LIVE-Qualcomm Mobile In-Capture Video Quality Database (LIVE-Q) [74], LIVE Video Quality Challenge Database (LIVE-VQC) [75], Konstanz Natural Video Database (KoNViD-1K) [76], and the resulting Compressed Video quality database with Semi-Automatic Ratings (CVSAR) [77].

The LIVE-Q database aims at in-capture distortions. This database consists of a total of 208 videos with a resolution of 1920×1080. These videos contain six common categories of capture distortion, and the videos are 15 seconds with 30 frames per seconds (fps). The LIVE-VQC database containing 585 videos of unique content, the resolution of the videos of this dataset is 320×240 and 1920×1080 at 30fps of 10 seconds duration. The KoNViD-1K is a large public video database collected from YFCC100M video database. It contains 1200 videos, each with the resolution of 960×540. The videos are 8 seconds in duration with frame rates of 24/25/30fps. The CVSAR database contains 130 original videos with resolution of 1920×1080 and 1280×720. For each original video, there are 3 different encoding modes and 4 different quantization parameters, a total of 1560 distorted videos. The videos are encoded at three predominant frame rates: 24fps, 25fps and 30fps.

In order to provide quantitative performance assessment for objective VQA, the performance assessment procedure of VQEG is adopted [74]. In this paper, we use Spearman rank order correlation coefficient (SRCC) and Pearson linear correlation coefficient (PLCC), which are commonly adopted to evaluate the performance of objective VQA methods. The closer the absolute values of SRCC and PLCC are to 1, the stronger linear correlation between the predicted scores of the model and the MOS values. This indicates a higher accuracy of the model and better overall performance [21].

The experimental process follows the same protocol in [12–14, 18, 34, 69]. The distorted videos from LIVE-Q, LIVE-VQC, KoNViD-1K, and CVSAR databases were partitioned into two sets: 80% for training and 20% for testing. To minimize the influence of randomness in STFR-BVQA, the training and testing procedures were repeated 10 times. The final result was determined by selecting the median score from the 10 repeated times, providing a representative and trustworthy measure of the effectiveness of STFR-BVQA.

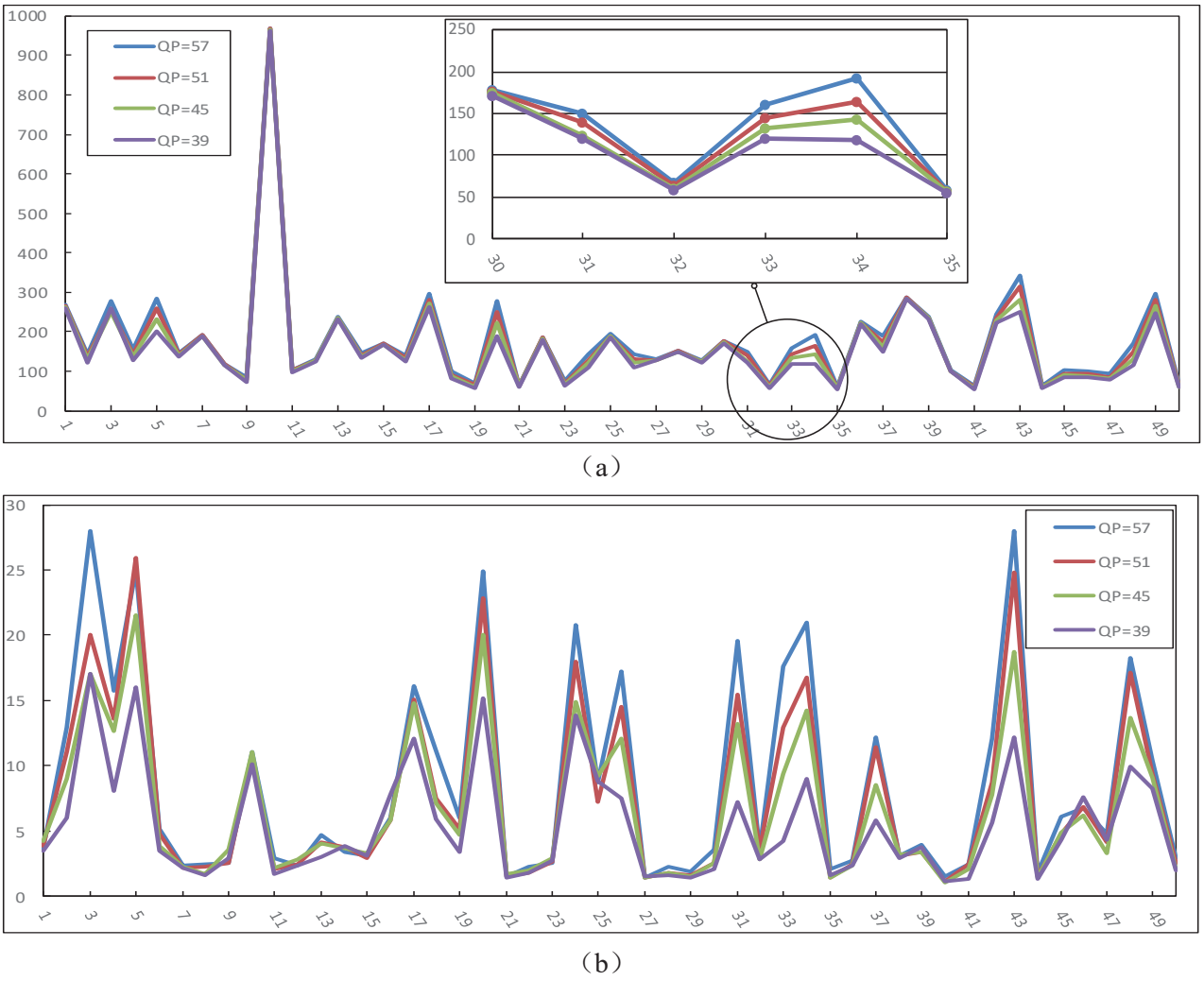

Figure 3: Curves of eigenvalues in video sequences with four different quantization parameters. (a)The horizontal axis represents the 50 **f_m**, and the vertical axis represents the corresponding **f_m** value. (b)The horizontal axis represents the 50 **f_s**, and the vertical axis represents the corresponding **f_s** value.

4.2. Performance on Individual Databases

We compared the STFR-BVQA with several classic BIQA methods (i.e., BRISQUE [38] and NIQE [39]) and other state-of-the-art BVQA methods (i.e., V-BLIINDS [10], VIIDEO [11], TLVQM [12], NSTSS [13], VIDEVAL [14], VSFA [16], CNN-TLVQM [18], LAGT-VQA [19], VISION [67], STEM [68], and VIQE [69]). These BVQA methods represent two different types of models: BVQA method based on manual feature extraction (i.e., V-BLIINDS, VIIDEO, TLVQM, NSTSS, VIDEVAL, VISION, STEM, and VIQE) and BVQA method based on deep learning (i.e., VSFA, CNN-TLVQM, and LAGT-VQA). Part of the source code and data provided by the respective authors.

Table 1 shows the results of SRCC and PLCC on the LIVE-Q, LIVE-VQC, KoNViD-1K, and CVSAR databases. As shown in Table 1, the proposed STFR-BVQA performs well across all databases. The SRCC results of BVQA-STFR have the top two performance of all the BVQA methods on LIVE-VQC, KoNViD-1K, and CVSAR, and ranks third in SRCC result on LIVE-Q. Additionally, the other regression models (NN and RFR) were used as the regression model instead of SVR. As can be seen from Table 1, SVR is selected as the regression model with slightly better performance, but there is little difference in their performance, which proves that the universality of our model in feature extraction has been verified.

4.3. Performance Across Databases

In addition, we evaluated the generalization capabilities of different BVQA methods. Training was first performed on the KoNViD-1K database and then evaluated on the LIVE-Q, LIVE-Q and CVSAR databases. As shown in Table 2, the STFR-BVQA achieves the top three performance in all BVQA methods. These databases cover a wide range of video content and distortion categories, which can pose challenges to BVQA methods based on manual feature extraction, limiting their ability to generalize. The BVQA methods based on deep learning

Table 1: SRCC AND PLCC RESULTS ON LIVE-Q, LIVE-VQC, KONVID-1K, AND CVSAR. THE TOP TWO RESULTS ARE HIGHLIGHTED IN BOLD

METHOD	LIVE-Q		LIVE-VQC		KoNViD-1K		CVSAR	
	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
2012-BRISQUE	0.545	0.549	0.633	0.595	0.354	0.657	0.601	0.548
2014-NIQE	0.481	0.467	0.546	0.632	0.411	0.378	0.435	0.420
2014-VBLIINDS	0.568	0.566	0.488	0.482	0.658	0.695	0.604	0.587
2016-VIIDEO	0.001	0.127	0.214	0.033	0.303	0.341	0.461	0.382
2019-VSFA	0.732	0.552	0.663	0.640	0.755	0.788	0.774	0.768
2019-TLVQM	**0.809**	0.769	0.432	0.450	0.764	0.760	0.692	0.684
2020-CNN-TLVQM	0.807	0.783	0.751	0.718	0.820	0.820	0.897	0.889
2020-NSTSS	0.628	0.589	0.655	0.647	0.653	0.641	0.754	0.732
2021-VIDEVAL	0.720	0.690	**0.752**	**0.751**	0.772	0.774	0.824	0.822
2022-LAGT-VQA	**0.852**	**0.842**	0.727	0.730	**0.864**	**0.841**	**0.924**	**0.900**
2022-VISION	0.584	0.547	0.689	0.676	0.632	0.598	—	—
2022-STEM	0.483	0.435	0.670	0.656	0.629	0.629	0.814	0.808
2023-VIQE	0.745	0.701	0.694	0.660	0.638	0.628	0.765	0.747
STFR-BVQA(NN)	0.771	0.769	0.800	0.791	0.811	0.809	0.884	0.869
STFR-BVQA(RFR)	0.755	0.744	0.790	0.789	0.814	0.814	0.845	0.850
STFR-BVQA	0.781	0.778	**0.805**	**0.801**	**0.826**	**0.822**	**0.912**	**0.906**

require extensive training on large-scale datasets to effectively extract quality-aware features using deep models. However, the feature extraction process of STFR-BVQA will not affect the judgment of video quality with the change of video content. Therefore, STFR-BVQA can be applied to a wider range of video content and distortion types. Moreover, during the evaluation process, it is only necessary to ensure that the block size of the test video sequence is consistent with that of the STFR training process. Therefore, the STFR does not require retraining during the cross-database evaluation process, and there is not need to resize the video sequences to the same size. This highlights the convenience and versatility of STFR-BVQA.

4.4. Statistical Significance Test

To estimated the statistical significance of STFR-BVQA, we performed a two sample t-test with a 95% confidence level. This test compared the SRCC values generated by various BVQA methods, including BRISQUE, NIQE, V-BLIINDS, VIIDEO, TLVQM, NSTSS, VIDEVAL, VSFA, CNN-TLVQM, LAGT-VQA, VISION, STEM, VIQE, and STFR-BVQA, across 100 iterations of experiments on the LIVE-VQC database. The results of the t-test are listed in Table 3 and Table 4, '1/-1/0' respectively represents the model in row is statistically better than/worse than/indistinguishable with the model in column with 95% confidence level [38, 51]. From Table 3 and Table 4, it becomes evident that STFR-BVQA is statistically better than the other BVQA methods.

4.5. Computational Efficiency

To compared the computational efficiency of STFR-BVQA with other BVQA methods, we selected two video sequences with different resolutions from the CVSAR database, in which the resolutions were 320 $\times$ 240 and 1920 $\times$ 1080. The length and frame rate of the video sequences were set to the same size, at 10 seconds and 30fps, respectively. From Table 5, it is evident that STFR-BVQA demonstrates the fastest computational speed compared to other BVQA methods across video sequences with different resolutions. Moreover, the higher the resolution of video sequences, the more obvious the calculation speed. It is worth noting that the time required for STFR training is 796.85 seconds. STFR only needs to be trained once, and the training time for STFR is not taken into account in this comparison.

At the same time, to further explore the performance results and computation time of different video segment lengths (10 frames, 30 frames, 60 frames, 100 frames, 150 frames, and 300 frames for the video segment). These video segment of different lengths were selected from the same video sequence. As shown in Table 6, the SRCC and PLCC results get better as the number of frames in the video segments increases. However, it should be noted that longer video segments lead to increased computation time.

In order to provide a more intuitive illustration, the line graph in Figure. 4 shows the SRCC and PLCC results for 30 different lengths of video segments (the number of video frames

Table 2: SRCC AND PLCC RESULTS ON LIVE-Q, LIVE-VQC, KONVID-1K, AND CVSAR. THE TOP TWO RESULTS ARE HIGHLIGHTED IN BOLD

TRAING-DATA	KoNViD-1K					
TEST-DATA	LIVE-Q		LIVE-VQC		CVSAR	
METRIC	PLCC	SRCC	PLCC	SRCC	PLCC	SRCC
2012-BRISQUE	0.156	0.132	0.488	0.419	0.523	0.514
2014-NIQE	0.321	0.258	0.508	0.500	0.647	0.655
2014-VBLIINDS	0.507	0.492	0.641	0.602	0.748	0.765
2016-VIIDEO	0.254	0.157	0.221	0.201	0.301	0.300
2019-VSFA	0.433	0.458	0.620	0.614	0.668	0.669
2019-TLVQM	0.388	0.387	**0.724**	0.689	0.789	0.768
2020-CNN-TLVQM	**0.640**	**0.633**	0.711	**0.693**	**0.824**	**0.830**
2020-NSTSS	0.532	0.512	0.541	0.540	0.665	0.641
2021-VIDEVAL	0.355	0.345	0.521	0.517	0.648	0.624
2022-LAGT-VQA	**0.668**	**0.651**	0.718	0.691	**0.842**	**0.831**
2022-VISION	0.512	0.510	0.621	0.617	—	—
2022-STEM	0.443	0.428	0.601	0.589	0.614	0.601
2023-VIQE	0.632	0.614	0.611	0.611	0.778	0.768
STFR-BVQA	0.638	0.622	**0.738**	**0.712**	0.814	0.801

Table 3: RESULTS OF THE TWO SAMPLE T-TEST PERFORMED BETWEEN SRCC VALUES OBTAINED BY DIFFERENT METHODS(BVQA METHODS BASED ON MANUAL FEATURE EXTRACTION)

LIVE-VQC	VBLIINDS	VIIDEO	TLVQM	NSTSS	VIDEVAL	VISION	STEM	VIQE	STFR-BVQA
2014-VBLIINDS	0	1	-1	-1	-1	-1	-1	-1	-1
2016-VIIDEO	-1	0	-1	-1	-1	-1	-1	-1	-1
2019-TLVQM	1	1	0	-1	-1	-1	-1	-1	-1
2020-NSTSS	1	1	1	0	-1	-1	-1	-1	-1
2021-VIDEVAL	1	1	1	1	0	1	1	1	-1
2022-VISION	1	1	1	1	-1	0	-1	-1	-1
2022-STEM	1	1	1	1	-1	1	0	0	-1
2023-VIQE	1	1	1	1	-1	1	0	0	-1
STFR-BVQA	1	1	1	1	1	1	1	1	0

ranges from 300 to 10, decrementing by 10) at a resolution of 1920×1080, as well as the computation time taken. The left side of the vertical axis shows the results of SRCC and PLCC, while the right side of the vertical axis shows the time required for calculation. The horizontal axis represents the number of video frames in the video segment. As the number of video frames increases, the performance of STFR-BVQA (PLCC and SRCC) increases, but so does the computation time. We also observed that the performance of STFR-BVQA gradually levels off as the number of video frames exceeds 100. Therefore, when the accuracy of video quality score calculation is not critical, it is possible to select part of the video sequence to calculate the quality score of the whole video sequence. The STFR-BVQA not only guarantees the calculation accuracy, but also speeds up the calculation speed. The selection of video segments can be either random or regular, and the results remain consistent.

Table 4: RESULTS OF THE TWO SAMPLE T-TEST PERFORMED BETWEEN SRCC VALUES OBTAINED BY DIFFERENT METHODS(BVQA METHOD BASED ON DEEP LEARNING)

LIVE-VQC	VSFA	CNN-TLVQM	LAGT-VQA	STFR-BVQA
2019-VSFA	0	-1	-1	-1
2020-CNN-TLVQM	1	0	0	-1
2022-LAGT-VQA	1	0	0	-1
STFR-BVQA	1	1	1	0

Table 5: THE TIME OF VIDEO SEQUENCES WITH DIFFERENT RESOLUTIONS IS CALCULATED IN DIFFERENT QUALITY ASSESSMENT METHODS.(RESOLUTION/FPS/SECONDS)

METHOD	RESOLUTION/FPS/SECONDS	
	360×240/30/10	1920×1080/30/10
2012-BRISQUE	15.1	103.2
2014-NIQE	17.3	234.6
2014-VBLIINDS	148.9	2024.9
2016-VIIDEO	52.8	683.0
2019-VSFA	102.6	1387.9
2019-TLVQM	21.4	256.5
2020-CNN-TLVQM	50.0	123.4
2020-NSTSS	45.8	263.8
2021-VIDEVAL	149.9	1798.6
2022-LAGT-VQA	532.1	1066.7
2022-VISION	42.1	134.5
2022-STEM	127.1	318.0
2023-VIQE	16.7	200.6
STFR-BVQA	14.3	19.1

Table 6: THE AVERAGE COMPUTATION TIME (SECONDS) FOR VIDEOS SELECTED FROM THE ORIGINAL DATABASES. (RESOLUTION/FPS/SECONDS)

Video Segment Length/Frames	PLCC	SRCC	Computation Time(s)
1920×1080/300(ALL)	0.849	0.848	19.1
1920×1080/150	0.844	0.840	18.6
1920×1080/100	0.842	0.837	15.4
1920×1080/60	0.834	0.833	8.8
1920×1080/30	0.831	0.819	4.5
1920×1080/10	0.810	0.817	1.9

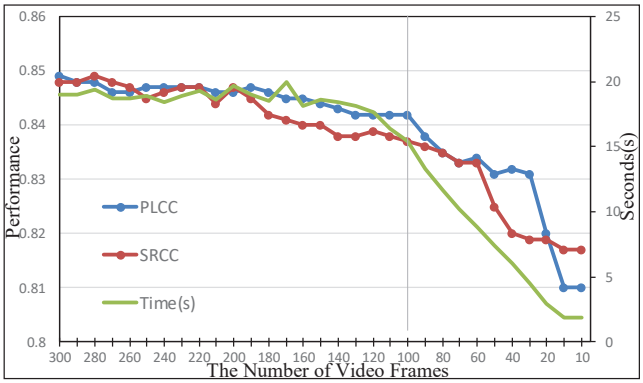

Figure 4: Effects of different video frame numbers on STFR-BVQA performance and computation time.

Table 7: AVERAGE SRCC AND PLCC RESULTS OF ABLATION EXPERIMENTS ACROSS TEN SESSIONS

Metric	PLCC	SRCC
f_m	0.614	0.608
f_r	0.415	0.301
f_s	0.547	0.564
f_m+f_r	0.594	0.610
f_m+f_s	0.801	0.789
f_r+f_s	0.698	0.647
f_m+f_r+f_s (all)	0.847	0.846

4.6. Ablation Study

In this subsection, we conducted ablation experiments to evaluate the contribution of key components in STFR-BVQA. STFR-BVQA contains three features, **f_m**, **f_r** and **f_s**. Table 7 lists the SRCC results of the STFR-BVQA feature subsets on KoNViD-1K database. We first selected **f_m** (**f_r** or **f_s**) as independent features for the experiment. However, the performance was unsatisfactory and hardly reflected significant improvement. These results indicate that only relying on spatial information, temporal information, or special video frame information is insufficient for the complex VQA task. Furthermore, when combining two of the three features, the performance was relatively worse compared to the complete STFR-BVQA. This highlights the importance of incorporating all three features in achieving the best overall performance. In summary, all the features in STFR-BVQA contribute to its overall performance, and the fusion of these features yields the optimal results.

4.7. Feature Extraction Model Replacement

In order to further verify the effectiveness of STFR in feature extraction for BVQA methods, we selected the feature extraction part of the existing trained deep network models (such as AlexNet [78], VGG [79] and GoogLeNet [80]) as the deep feature extractor to replace STFR, and conducted a comparison test

Table 8: COMPARING THE AVERAGE SRCC AND PLCC RESULTS, ALONG WITH CALCULATION TIME,BY REPLACING DIFFERENT FEATURE EXTRACTORS IN STFR-BVQA ON THE LIVE-VQC DATABASE ACROSS TEN SESSIONS

METHOD	PLCC	SRCC	Computation Time(s)
AlexNet-BVQA	0.529	0.494	36.5
VGG16-BVQA	0.578	0.541	65.4
VGG19-BVQA	0.665	0.608	79.1
GoogLeNet-BVQA	0.577	0.528	47.8
STFR-BVQA	**0.805**	**0.801**	**19.5**

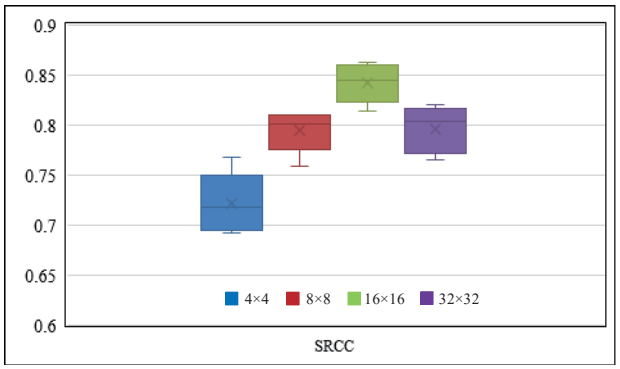

Figure 5: Box plot of the effect of video block size on performance during STFR training.

with STFR-BVQA. As shown in Table 8, which lists the SRCC and PLCC results in the LIVE-VQC database and demonstrates the time required to compute a video sequence with a resolution of 1920×1080 at 30fps for a duration of 10 seconds. From Table 8, STFR-BVQA outperforms BVQA based on the deep feature extractor both in terms of performance and calculation time.The reasons for this phenomenon are as follows: 1) Although the deep feature extractor can extract more advanced features, these network models are based on visual tasks and are not well-suited for simulating the HVS in quality assessment; 2) The existing deep feature extractors are image-oriented and struggle to analyze and extract the relevant feature information from videos; 3) The deep feature extractor, with its abundance of network parameters and layers, it is more complex than STFR; 4) Most importantly, STFR is obtained through the analysis of video characteristics and training, and has universality. It can be used for videos of different scenes after training once.

4.8. Analysis of Parameters

There is a crucial parameter, denoted as S, in STFR training, which represents the size of the segmented video block. In order to investigate the impact of S on performance, four values of S (4,8,16,32) were selected for this test.

Figure. 5 shows the effect of S size on performance in the KoNViD-1K database. From Figure. 5, it is evident that for different values of S, STFR-BVQA achieves the best performance when $S = 16$.In addition to the experiment on KoNViD-1K database, we also reached the same conclusion with experiments in the LIVE-Q, LIVE-VQC, and CVSAR databases. After many comparisons, $S = 16$ was finally selected as the optimal value.

5. Conclusions

In this paper, we proposed a BVQA method called STFR-BVQA. Our approach utilizes an independent feature extraction process, where STFR is trained to extract spatio-temporal features from video sequences. It is worth mentioning that STFR is universal and can be directly applied to extract spatio-temporal features from different videos with just one training session, and the training process of STFR only requires one video sequence. During the quality assessment process, STFR is firstly used to automatically extract spatio-temporal features from videos through multiplication. Then, according to the similarity and difference of data distribution, representative spatio-temporal features of the video sequence are computed. Finally, these spatio-temporal features are combined into feature vectors, and a regression model, SVR, is used to map the video quality scores.The experimental results conducted on the LIVE-Q, LIVE-VQC, KoNViD-1K and CVSAR databases show that STFR-BVQA is superior to all BVQA methods based on manual feature extraction in terms of computational accuracy. While the improvement in performance compared to deep learning-based BVQA methods may not be significant, STFR-BVQA exhibits a obvious advantage in computational speed.

Acknowledgement

This work is supported by the National Natural Science Foundation of China (Grant No. 62271336 and Grant No. 62211530110).

References

[1] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, L. K. Cormack, Study of Subjective and Objective Quality Assessment of Video, IEEE Transactions on Image Processing 19 (6) (2010) 1427–1441. doi:10.1109/TIP.2010.2042111.
URL http://ieeexplore.ieee.org/document/5404314/

[2] X. Dong, H. Zhou, Texture synthesis quality assessment using perceptual texture similarity, Knowledge-Based Systems 194 (2020) 105591. doi:10.1016/j.knosys.2020.105591.
URL https://linkinghub.elsevier.com/retrieve/pii/S095070512030068X

[3] M. Narwaria, Weisi Lin, Anmin Liu, Low-Complexity Video Quality Assessment Using Temporal Quality Variations, IEEE Transactions on Multimedia 14 (3) (2012) 525–535. doi:10.1109/TMM.2012.2190589.
URL http://ieeexplore.ieee.org/document/6168273/

[4] S. Chikkerur, V. Sundaram, M. Reisslein, L. J. Karam, Objective Video Quality Assessment Methods: A Classification, Review, and Performance Comparison, IEEE Transactions on Broadcasting 57 (2) (2011) 165–182. doi:10.1109/TBC.2011.2104671.
URL http://ieeexplore.ieee.org/document/5710601/

[5] J. Yang, Y. Zhu, C. Ma, W. Lu, Q. Meng, Stereoscopic video quality assessment based on 3D convolutional neural networks, Neurocomputing 309 (2018) 83–93. doi:10.1016/j.neucom.2018.04.072.
URL https://linkinghub.elsevier.com/retrieve/pii/S0925231218305290

[6] C. G. Bampis, Z. Li, A. C. Bovik, Spatiotemporal Feature Integration and Model Fusion for Full Reference Video Quality Assessment, IEEE Transactions on Circuits and Systems for Video Technology 29 (8) (2019) 2256–2270. doi:10.1109/TCSVT.2018.2868262.
URL https://ieeexplore.ieee.org/document/8453837/

[7] Y. Zhang, L. He, W. Lu, J. Li, X. Gao, Video quality assessment with dense features and ranking pooling, Neurocomputing 457 (2021) 242–253. doi:10.1016/j.neucom.2021.06.026.
URL https://linkinghub.elsevier.com/retrieve/pii/S0925231221009309

[8] L. K. Choi, A. C. Bovik, Video quality assessment accounting for temporal visual masking of local flicker, Signal Processing: Image Communication 67 (2018) 182–198. doi:10.1016/j.image.2018.06.009.
URL https://linkinghub.elsevier.com/retrieve/pii/S0923596518306374

[9] L. Ma, S. Li, K. N. Ngan, Reduced-Reference Video Quality Assessment of Compressed Video Sequences, IEEE Transactions on Circuits and Systems for Video Technology 22 (10) (2012) 1441–1456. doi:10.1109/TCSVT.2012.2202049.
URL http://ieeexplore.ieee.org/document/6209406/

[10] M. A. Saad, A. C. Bovik, C. Charrier, Blind Prediction of Natural Video Quality, IEEE Transactions on Image Processing 23 (3) (2014) 1352–1365. doi:10.1109/TIP.2014.2299154.
URL http://ieeexplore.ieee.org/document/6705673/

[11] A. Mittal, M. A. Saad, A. C. Bovik, A Completely Blind Video Integrity Oracle, IEEE Transactions on Image Processing 25 (1) (2016) 289–300. doi:10.1109/TIP.2015.2502725.
URL http://ieeexplore.ieee.org/document/7332944/

[12] J. Korhonen, Two-Level Approach for No-Reference Consumer Video Quality Assessment, IEEE Transactions on Image Processing 28 (12) (2019) 5923–5938. doi:10.1109/TIP.2019.2923051.
URL https://ieeexplore.ieee.org/document/8742797/

[13] S. V. Reddy Dendi, S. S. Channappayya, No-Reference Video Quality Assessment Using Natural Spatiotemporal Scene Statistics, IEEE Transactions on Image Processing 29 (2020) 5612–5624. doi:10.1109/TIP.2020.2984879.
URL https://ieeexplore.ieee.org/document/9059006/

[14] Z. Tu, Y. Wang, N. Birkbeck, B. Adsumilli, A. C. Bovik, UGC-VQA: Benchmarking Blind Video Quality Assessment for User Generated Content, IEEE Transactions on Image Processing 30 (2021) 4449–4464. doi:10.1109/TIP.2021.3072221.
URL https://ieeexplore.ieee.org/document/9405420/

[15] Y. Li, L.-M. Po, C.-H. Cheung, X. Xu, L. Feng, F. Yuan, K.-W. Cheung, No-Reference Video Quality Assessment With 3D Shearlet Transform and Convolutional Neural Networks, IEEE Transactions on Circuits and Systems for Video Technology 26 (6) (2016) 1044–1057. doi:10.1109/TCSVT.2015.2430711.
URL http://ieeexplore.ieee.org/document/7103036/

[16] D. Li, T. Jiang, M. Jiang, Quality Assessment of In-the-Wild Videos, in: Proceedings of the 27th ACM International Conference on Multimedia, ACM, Nice France, 2019, pp. 2351–2359. doi:10.1145/3343031.3351028.
URL https://dl.acm.org/doi/10.1145/3343031.3351028

[17] P. Chen, L. Li, L. Ma, J. Wu, G. Shi, RIRNet: Recurrent-In-Recurrent Network for Video Quality Assessment, in: Proceedings of the 28th ACM International Conference on Multimedia, ACM, Seattle WA USA, 2020, pp. 834–842. doi:10.1145/3394171.3413717.
URL https://dl.acm.org/doi/10.1145/3394171.3413717

[18] J. Korhonen, Y. Su, J. You, Blind Natural Video Quality Prediction via Statistical Temporal Features and Deep Spatial Features, in: Proceedings of the 28th ACM International Conference on Multimedia, ACM, Seattle WA USA, 2020, pp. 3311–3319. doi:10.1145/3394171.3413845.
URL https://dl.acm.org/doi/10.1145/3394171.3413845

[19] J. You, Y. Lin, Efficient Transformer with Locally Shared Attention for Video Quality Assessment, in: 2022 IEEE International Conference on Image Processing (ICIP), IEEE, Bordeaux, France, 2022, pp. 356–360. doi:10.1109/ICIP46576.2022.9898025.
URL https://ieeexplore.ieee.org/document/9898025/

[20] W. Shen, M. Zhou, X. Liao, W. Jia, T. Xiang, B. Fang, Z. Shang, An End-to-End No-Reference Video Quality Assessment Method With Hierarchical Spatiotemporal Feature Representation, IEEE Transactions on Broadcasting 68 (3) (2022) 651–660. doi:10.1109/TBC.2022.3164332.
URL https://ieeexplore.ieee.org/document/9754977/

[21] H.-W. Chang, X.-D. Bi, C. Kai, Blind Image Quality Assessment by Visual Neuron Matrix, IEEE Signal Processing Letters 28 (2021) 1803–1807. doi:10.1109/LSP.2021.3106579.
URL https://ieeexplore.ieee.org/document/9520276/

[22] L. He, W. Lu, C. Jia, L. Hao, Video quality assessment by compact representation of energy in 3D-DCT domain, Neurocomputing 269 (2017) 108–116. doi:10.1016/j.neucom.2016.08.143.
URL https://linkinghub.elsevier.com/retrieve/pii/S0925231217309682

[23] Z. Wang, A. Bovik, H. Sheikh, E. Simoncelli, Image Quality Assessment: From Error Visibility to Structural Similarity, IEEE Transactions on Image Processing 13 (4) (2004) 600–612. doi:10.1109/TIP.2003.819861.
URL http://ieeexplore.ieee.org/document/1284395/

[24] Sameeulla Khan Md, S. S. Channappayya, Multiscale-SSIM index based stereoscopic image quality assessment, in: 2016 Twenty Second National Conference on Communication (NCC), IEEE, Guwahati, India, 2016, pp. 1–5. doi:10.1109/NCC.2016.7561082.
URL http://ieeexplore.ieee.org/document/7561082/

[25] Y. Jin, A. Patney, R. Webb, A. C. Bovik, FOVQA: Blind Foveated Video Quality Assessment, IEEE Transactions on Image Processing 31 (2022) 4571–4584. doi:10.1109/TIP.2022.3185738.
URL https://ieeexplore.ieee.org/document/9810800/

[26] M. Pinson, S. Wolf, A New Standardized Method for Objectively Measuring Video Quality, IEEE Transactions on Broadcasting 50 (3) (2004) 312–322. doi:10.1109/TBC.2004.834028.
URL http://ieeexplore.ieee.org/document/1337089/

[27] X. Li, Q. Guo, X. Lu, Spatiotemporal Statistics for Video Quality Assessment, IEEE Transactions on Image Processing 25 (7) (2016) 3329–3342. doi:10.1109/TIP.2016.2568752.
URL http://ieeexplore.ieee.org/document/7469872/

[28] X. Guan, F. Li, Y. Zhang, P. C. Cosman, End-to-End Blind Video Quality Assessment Based on Visual and Memory Attention Modeling, IEEE Transactions on Multimedia (2022) 1–16doi:10.1109/TMM.2022.3189251.
URL https://ieeexplore.ieee.org/document/9817621/

[29] Y. Liu, J. Wu, L. Li, W. Dong, J. Zhang, G. Shi, Spatiotemporal Representation Learning for Blind Video Quality Assessment, IEEE Transactions on Circuits and Systems for Video Technology 32 (6) (2022) 3500–3513. doi:10.1109/TCSVT.2021.3114509.
URL https://ieeexplore.ieee.org/document/9546044/

[30] Y. Zhang, X. Gao, L. He, W. Lu, R. He, Blind Video Quality Assessment With Weakly Supervised Learning and Resampling Strategy, IEEE Transactions on Circuits and Systems for Video Technology 29 (8) (2019) 2244–2255. doi:10.1109/TCSVT.2018.2868063.
URL https://ieeexplore.ieee.org/document/8453041/

[31] W. Liu, Z. Duanmu, Z. Wang, End-to-End Blind Quality Assessment of Compressed Videos Using Deep Neural Networks, in: Proceedings of the 26th ACM international conference on Multimedia, ACM, Seoul Republic of Korea, 2018, pp. 546–554. doi:10.1145/3240508.3240643.
URL https://dl.acm.org/doi/10.1145/3240508.3240643

[32] M. T. Vega, D. C. Mocanu, J. Famaey, S. Stavrou, A. Liotta, Deep Learning for Quality Assessment in Live Video Streaming, IEEE Signal Processing Letters 24 (6) (2017) 736–740. doi:10.1109/LSP.2017.2691160.
URL http://ieeexplore.ieee.org/document/7892928/

[33] W. Y. Akamine, P. G. Freitas, M. C. Farias, A framework for computationally efficient video quality assessment, Signal Processing: Image Communication 70 (2019) 57–67. doi:10.1016/j.image.2018.09.009.
URL https://linkinghub.elsevier.com/retrieve/pii/S0923596518308749

[34] M. Zhou, S. Lang, T. Zhang, X. Liao, Z. Shang, T. Xiang, B. Fang, Attentional Feature Fusion for End-to-End Blind Image Quality Assessment, IEEE Transactions on Broadcasting 69 (1) (2023) 144–152. doi:10.1109/TBC.2022.3204235.

URL https://ieeexplore.ieee.org/document/9893864/
[35] H.-w. Chang, Q.-w. Zhang, Q.-g. Wu, Y. Gan, Perceptual image quality assessment by independent feature detector, Neurocomputing 151 (2015) 1142–1152. doi:10.1016/j.neucom.2014.04.081.
URL https://linkinghub.elsevier.com/retrieve/pii/S0925231214013344
[36] A. K. Moorthy, A. C. Bovik, Blind Image Quality Assessment: From Natural Scene Statistics to Perceptual Quality, IEEE Transactions on Image Processing 20 (12) (2011) 3350–3364. doi:10.1109/TIP.2011.2147325.
URL http://ieeexplore.ieee.org/document/5756237/
[37] M. A. Saad, A. C. Bovik, C. Charrier, Blind Image Quality Assessment: A Natural Scene Statistics Approach in the DCT Domain, IEEE Transactions on Image Processing 21 (8) (2012) 3339–3352. doi:10.1109/TIP.2012.2191563.
URL http://ieeexplore.ieee.org/document/6172573/
[38] A. Mittal, A. K. Moorthy, A. C. Bovik, No-Reference Image Quality Assessment in the Spatial Domain, IEEE Transactions on Image Processing 21 (12) (2012) 4695–4708. doi:10.1109/TIP.2012.2214050.
URL http://ieeexplore.ieee.org/document/6272356/
[39] A. Mittal, R. Soundararajan, A. C. Bovik, Making a Completely Blind Image Quality Analyzer, IEEE Signal Processing Letters 20 (3) (2013) 209–212. doi:10.1109/LSP.2012.2227726.
URL http://ieeexplore.ieee.org/document/6353522/
[40] L. Zhang, L. Zhang, A. C. Bovik, A Feature-Enriched Completely Blind Image Quality Evaluator, IEEE Transactions on Image Processing 24 (8) (2015) 2579–2591. doi:10.1109/TIP.2015.2426416.
URL https://ieeexplore.ieee.org/document/7094273/
[41] P. C. Madhusudana, N. Birkbeck, Y. Wang, B. Adsumilli, A. C. Bovik, Image Quality Assessment Using Contrastive Learning, IEEE Transactions on Image Processing 31 (2022) 4149–4161. doi:10.1109/TIP.2022.3181496.
URL https://ieeexplore.ieee.org/document/9796010/
[42] T.-H. Chan, K. Jia, S. Gao, J. Lu, Z. Zeng, Y. Ma, PCANet: A Simple Deep Learning Baseline for Image Classification?, IEEE Transactions on Image Processing 24 (12) (2015) 5017–5032. doi:10.1109/TIP.2015.2475625.
URL http://ieeexplore.ieee.org/document/7234886/
[43] Y. Bengio, A. Courville, P. Vincent, Representation Learning: A Review and New Perspectives, IEEE Transactions on Pattern Analysis and Machine Intelligence 35 (8) (2013) 1798–1828. doi:10.1109/TPAMI.2013.50.
URL http://ieeexplore.ieee.org/document/6472238/
[44] J. Xiang, M. Yu, G. Jiang, X. Zhang, H. Xu, Blind light field image quality assessment with tensor color domain and 3D shearlet transform, Signal Processing 211 (2023) 109083. doi:10.1016/j.sigpro.2023.109083.
URL https://linkinghub.elsevier.com/retrieve/pii/S0165168423001573
[45] M. T. Rasheed, D. Shi, H. Khan, A comprehensive experiment-based review of low-light image enhancement methods and benchmarking low-light image quality assessment, Signal Processing 204 (2023) 108821. doi:10.1016/j.sigpro.2022.108821.
URL https://linkinghub.elsevier.com/retrieve/pii/S0165168422003607
[46] K. Zhang, D. Zhu, J. Li, X. Gao, F. Gao, J. Lu, Learning stacking regression for no-reference super-resolution image quality assessment, Signal Processing 178 (2021) 107771. doi:10.1016/j.sigpro.2020.107771.
URL https://linkinghub.elsevier.com/retrieve/pii/S0165168420303145
[47] H. Zhu, L. Li, J. Wu, W. Dong, G. Shi, MetaIQA: Deep Meta-Learning for No-Reference Image Quality Assessment, in: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Seattle, WA, USA, 2020, pp. 14131–14140. doi:10.1109/CVPR42600.2020.01415.
URL https://ieeexplore.ieee.org/document/9156932/
[48] L. Kang, P. Ye, Y. Li, D. Doermann, Convolutional Neural Networks for No-Reference Image Quality Assessment, in: 2014 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, Columbus, OH, USA, 2014, pp. 1733–1740. doi:10.1109/CVPR.2014.224.
URL https://ieeexplore.ieee.org/document/6909620
[49] K. Ma, W. Liu, K. Zhang, Z. Duanmu, Z. Wang, W. Zuo, End-to-End Blind Image Quality Assessment Using Deep Neural Networks, IEEE Transactions on Image Processing 27 (3) (2018) 1202–1213. doi:10.1109/TIP.2017.2774045.
URL http://ieeexplore.ieee.org/document/8110690/
[50] W. Zhang, K. Ma, J. Yan, D. Deng, Z. Wang, Blind Image Quality Assessment Using a Deep Bilinear Convolutional Neural Network, IEEE Transactions on Circuits and Systems for Video Technology 30 (1) (2020) 36–47. doi:10.1109/TCSVT.2018.2886771.
URL https://ieeexplore.ieee.org/document/8576582/
[51] J. Ma, J. Wu, L. Li, W. Dong, X. Xie, G. Shi, W. Lin, Blind Image Quality Assessment With Active Inference, IEEE Transactions on Image Processing 30 (2021) 3650–3663. doi:10.1109/TIP.2021.3064195.
URL https://ieeexplore.ieee.org/document/9376644/
[52] Z. Pan, F. Yuan, J. Lei, Y. Fang, X. Shao, S. Kwong, VCRNet: Visual Compensation Restoration Network for No-Reference Image Quality Assessment, IEEE Transactions on Image Processing 31 (2022) 1613–1627. doi:10.1109/TIP.2022.3144892.
URL https://ieeexplore.ieee.org/document/9694502/
[53] M. Liu, J. Huang, D. Zeng, X. Ding, J. Paisley, A Multiscale Approach to Deep Blind Image Quality Assessment, IEEE Transactions on Image Processing 32 (2023) 1656–1667. doi:10.1109/TIP.2023.3245991.
URL https://ieeexplore.ieee.org/document/10056825/
[54] S. Sun, T. Yu, J. Xu, W. Zhou, Z. Chen, GraphIQA: Learning Distortion Graph Representations for Blind Image Quality Assessment, IEEE Transactions on Multimedia (2022) 1–1doi:10.1109/TMM.2022.3152942.
URL https://ieeexplore.ieee.org/document/9720138/
[55] A. Li, J. Wu, S. Tian, L. Li, W. Dong, G. Shi, Blind image quality assessment based on progressive multi-task learning, Neurocomputing 500 (2022) 307–318. doi:10.1016/j.neucom.2022.05.043.
URL https://linkinghub.elsevier.com/retrieve/pii/S0925231222006087
[56] T. Song, L. Li, P. Chen, H. Liu, J. Qian, Blind Image Quality Assessment for Authentic Distortions by Intermediary Enhancement and Iterative Training, IEEE Transactions on Circuits and Systems for Video Technology 32 (11) (2022) 7592–7604. doi:10.1109/TCSVT.2022.3179744.
URL https://ieeexplore.ieee.org/document/9786803/
[57] R. Gao, Z. Huang, S. Liu, QL-IQA: Learning distance distribution from quality levels for blind image quality assessment, Signal Processing: Image Communication 101 (2022) 116576. doi:10.1016/j.image.2021.116576.
URL https://linkinghub.elsevier.com/retrieve/pii/S092359652100285X
[58] W. Zhang, D. Li, C. Ma, G. Zhai, X. Yang, K. Ma, Continual Learning for Blind Image Quality Assessment, IEEE Transactions on Pattern Analysis and Machine Intelligence (2022) 1–1doi:10.1109/TPAMI.2022.3178874.
URL https://ieeexplore.ieee.org/document/9784904/
[59] T. Li, X. Min, H. Zhao, G. Zhai, Y. Xu, W. Zhang, Subjective and Objective Quality Assessment of Compressed Screen Content Videos, IEEE Transactions on Broadcasting 67 (2) (2021) 438–449. doi:10.1109/TBC.2020.3028335.
URL https://ieeexplore.ieee.org/document/9234440/
[60] L. Liu, T. Wang, H. Huang, A. C. Bovik, Video quality assessment using spacetime slice mappings, Signal Processing: Image Communication 82 (2020) 115749. doi:10.1016/j.image.2019.115749.
URL https://linkinghub.elsevier.com/retrieve/pii/S092359651931077X
[61] H. Sheikh, A. Bovik, Image information and visual quality, IEEE Transactions on Image Processing 15 (2) (2006) 430–444. doi:10.1109/TIP.2005.859378.
URL https://ieeexplore.ieee.org/document/1576816
[62] Lin Zhang, Lei Zhang, Xuanqin Mou, D. Zhang, FSIM: A Feature Similarity Index for Image Quality Assessment, IEEE Transactions on Image Processing 20 (8) (2011) 2378–2386. doi:10.1109/TIP.2011.2109730.
URL http://ieeexplore.ieee.org/document/5705575/
[63] W. Xue, L. Zhang, X. Mou, A. C. Bovik, Gradient Magnitude Similarity Deviation: A Highly Efficient Perceptual Image Quality Index, IEEE Transactions on Image Processing 23 (2) (2014) 684–695. doi:10.1109/TIP.2013.2293423.

URL http://ieeexplore.ieee.org/document/6678238/

[64] Y. Fang, X. Sui, J. Yan, Y. Zuo, J. Wang, Z. Li, Asymmetrically distorted 3D video quality assessment: From the motion variation to perceived quality, Signal Processing 183 (2021) 108031. doi:10.1016/j.sigpro.2021.108031.
URL https://linkinghub.elsevier.com/retrieve/pii/S0165168421000700

[65] T. Sun, S. Ding, W. Chen, Blind video quality assessment based on multilevel video perception, Signal Processing: Image Communication 99 (2021) 116485. doi:10.1016/j.image.2021.116485.
URL https://linkinghub.elsevier.com/retrieve/pii/S0923596521002368

[66] S. Jiang, Q. Sang, Z. Hu, L. Liu, Self-Supervised Representation Learning for Video Quality Assessment, IEEE Transactions on Broadcasting 69 (1) (2023) 118–129. doi:10.1109/TBC.2022.3197904.
URL https://ieeexplore.ieee.org/document/9861385/

[67] S. Mitra, R. Soundararajan, Multiview Contrastive Learning for Completely Blind Video Quality Assessment of User Generated Content, in: Proceedings of the 30th ACM International Conference on Multimedia, ACM, Lisboa Portugal, 2022, pp. 1914–1924. doi:10.1145/3503161.3548064.
URL https://dl.acm.org/doi/10.1145/3503161.3548064

[68] P. Kancharla, S. S. Channappayya, Completely Blind Quality Assessment of User Generated Video Content, IEEE Transactions on Image Processing 31 (2022) 263–274. doi:10.1109/TIP.2021.3130541.
URL https://ieeexplore.ieee.org/document/9633248/

[69] Q. Zheng, Z. Tu, X. Zeng, A. C. Bovik, Y. Fan, A Completely Blind Video Quality Evaluator, IEEE Signal Processing Letters 29 (2022) 2228–2232. doi:10.1109/LSP.2022.3215311.
URL https://ieeexplore.ieee.org/document/9921340/

[70] A. H. Bakhtiari, A. Mansouri, No-Reference Video Quality Assessment by Deep Feature Maps Relations, in: 2022 12th International Conference on Computer and Knowledge Engineering (ICCKE), IEEE, Mashhad, Iran, Islamic Republic of, 2022, pp. 326–331. doi:10.1109/ICCKE57176.2022.9960078.
URL https://ieeexplore.ieee.org/document/9960078/

[71] X. Peng, D. Xu, Projection support vector regression algorithms for data regression, Knowledge-Based Systems 112 (2016) 54–66. doi:10.1016/j.knosys.2016.08.030.
URL https://linkinghub.elsevier.com/retrieve/pii/S0950705116303070

[72] X. Guo, Y. Qian, L. Li, A. Asano, Assessment model for perceived visual complexity of painting images, Knowledge-Based Systems 159 (2018) 110–119. doi:10.1016/j.knosys.2018.06.006.
URL https://linkinghub.elsevier.com/retrieve/pii/S0950705118303150

[73] M. Heydari, P. Cheraaqee, A. Mansouri, A. Mahmoudi-Aznaveh, A low complexity wavelet-based blind image quality evaluator, Signal Processing: Image Communication 74 (2019) 280–288. doi:10.1016/j.image.2018.12.016.
URL https://linkinghub.elsevier.com/retrieve/pii/S0923596518311809

[74] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, K.-C. Yang, In-Capture Mobile Video Distortions: A Study of Subjective Behavior and Objective Algorithms, IEEE Transactions on Circuits and Systems for Video Technology 28 (9) (2018) 2061–2077. doi:10.1109/TCSVT.2017.2707479.
URL https://ieeexplore.ieee.org/document/7932975/

[75] Z. Sinno, A. C. Bovik, Large-Scale Study of Perceptual Video Quality, IEEE Transactions on Image Processing 28 (2) (2019) 612–627. doi:10.1109/TIP.2018.2869673.
URL https://ieeexplore.ieee.org/document/8463581/

[76] V. Hosu, F. Hahn, M. Jenadeleh, H. Lin, H. Men, T. Sziranyi, S. Li, D. Saupe, The Konstanz natural video database (KoNViD-1k), in: 2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX), IEEE, Erfurt, Germany, 2017, pp. 1–6. doi:10.1109/QoMEX.2017.7965673.
URL http://ieeexplore.ieee.org/document/7965673/

[77] L. Lin, Z. Wang, J. He, W. Chen, Y. Xu, T. Zhao, Deep Quality Assessment of Compressed Videos: A Subjective and Objective Study, IEEE Transactions on Circuits and Systems for Video Technology (2022) 1–1doi:10.1109/TCSVT.2022.3227039.
URL https://ieeexplore.ieee.org/document/9970717/

[78] A. Krizhevsky, I. Sutskever, G. E. Hinton, ImageNet classification with deep convolutional neural networks, Communications of the ACM 60 (6) (2017) 84–90. doi:10.1145/3065386.
URL https://dl.acm.org/doi/10.1145/3065386

[79] K. Simonyan, A. Zisserman, Very Deep Convolutional Networks for Large-Scale Image RecognitionArXiv:1409.1556 [cs] (Apr. 2015).
URL http://arxiv.org/abs/1409.1556

[80] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, Boston, MA, USA, 2015, pp. 1–9. doi:10.1109/CVPR.2015.7298594.
URL http://ieeexplore.ieee.org/document/7298594/