Deep Quality Assessment of Compressed Videos: A Subjective and Objective Study

Liqun Lin®, Member, IEEE, Zheng Wang®, Jiachen He, Weiling Chen®, Member, IEEE, Yiwen Xu®, Member, IEEE, and Tiesong Zhao, Senior Member, IEEE

Abstract—Video quality assessment is critical in optimizing video coding techniques. However, the state-of-the-art methods have limited performance, which is largely due to the lack of large-scale subjective databases for training. In this work, a semi-automatic labeling method is adopted to build a large-scale compressed video quality database, which allows us to label a large number of compressed videos with manageable human workload. The resulting Compressed Video quality database with Semi-Automatic Ratings (CVSAR), so far the largest of compressed video quality database. We train a no-reference compressed video quality assessment model with a 3D CNN for SpatioTemporal Feature Extraction and Evaluation (STTFEE). Experimental results demonstrate that the proposed method outperforms state-of-the-art metrics and achieves promising generalization performance in cross-database tests. The CVSAR database has been made publicly available. It can be accessed at https://github.com/Rocknroll194/CVSAR.

Index Terms—Video quality assessment, semi-auto rating, compressed video, deep network.

I. INTRODUCTION

The perceptual quality of a compressed video is an extremely important indicator in video coding. With the development of High Definition (HD)/Ultra HD (UHD) and 3D/360-degree videos, the video traffic shows an explosive growth trend [1]. To meet such a demand, it is necessary to improve network bandwidth and maximize video quality under a limited bitrate or bandwidth constraint. Lossy video codecs have been developed to address this issue, but the visual qualities of compressed videos are yet to be evaluated [2].

The perceptual quality of a compressed video is an extremely important indicator in video coding. With the development of High Definition (HD)/Ultra HD (UHD) and 3D/360-degree videos, the video traffic shows an explosive growth trend [1]. To meet such a demand, it is necessary to improve network bandwidth and maximize video quality under a limited bitrate or bandwidth constraint. Lossy video codecs have been developed to address this issue, but the visual qualities of compressed videos are yet to be evaluated [2].

Manuscript received 9 October 2022; revised 26 November 2022; accepted 27 November 2022. Date of publication 6 December 2022; date of current version 6 June 2023. This work was supported in part by the Natural Science Foundation of China under Grant 62171134 and Grant 61901119, in part by the Natural Science Foundation of Fujian Province under Grant 2022J02015, and in part by the Fujian Provincial Education Department under Grant JAT200024. This article was recommended by Associate Editor B. Zatt. (Corresponding author: Yiwen Xu.)

Liqun Lin and Tiesong Zhao are with the Fujian Key Laboratory for Intelligent Processing and Wireless Transmission of Media Information, College of Physics and Information Engineering, Fuzhou University, Fuzhou, Fujian 350116, China, and also with the Fujian Science and Technology Innovation Laboratory for Optoelectronic Information of China, Fuzhou, Fujian 350108, China (e-mail: Lin_lq@fzu.edu.cn; t.zhao@fzu.edu.cn).

Zheng Wang, Jiachen He, Weiling Chen, and Yiwen Xu, are with the Fujian Key Laboratory for Intelligent Processing and Wireless Transmission of Media Information, College of Physics and Information Engineering, Fuzhou University, Fuzhou, Fujian 350116, China (e-mail: n191120078@fzu.edu.cn; 067900190@fzu.edu.cn; weiling.chen@fzu.edu.cn; xu_yiwen@fzu.edu.cn).

Color versions of one or more figures in this article are available at https://doi.org/10.1109/TCSVT.2022.3227039.

Digital Object Identifier 10.1109/TCSVT.2022.3227039

Therefore, a reliable Video Quality Assessment (VQA) method for compressed videos is critical as both a performance indicator and a guidance for further improvement. Peak Signal-to-Noise Ratio (PSNR) [3] and Structural SIMilarity (SSIM) index [4] are the mainstream quality assessment methods. However, human visual characteristics and temporal characteristics between video frames are not considered, the perceptual quality of videos cannot be accurately expressed.

Existing VQA methods include subjective VQA methods and objective VQA models. Subjective VQA methods are subjective tests by human observers, which is the most reliable evaluation method because the user is the ultimate viewer of videos. However, this method is time-consuming and impractical. Instead, it is often used in the construction of various quality evaluation databases to provide a reliable reference for objective VQA models. Objective VQA models are guided by subjective scores and predict the quality of videos by automatic algorithms. Due to the convenience and low cost, objective VQA models are widely utilized.

Conventionally, objective VQA models utilize signal differences or hand-crafted image features to model compressed video quality. Earlier methods include the above mentioned PSNR and SSIM. Later, the compressed videos are usually evaluated as a simple composition of independent images, thus the Image Quality Assessment (IQA) methods have been widely applied [5], [6], [7]. To further exploit the temporal characteristics of video frames, the researchers have also utilized the temporal dependency or motion information in VQA of compressed videos [8], [9].

Recent researches have observed high-level features, e.g. compression artifacts [10], [11], [12], greatly influence the compressed video quality. However, it is hard to model these artifacts with hand-crafted features. As a result, Convolutional Neural Networks (CNNs) have been utilized to detect compression artifacts for VQA [13]. Besides, the deep learning has also shown its strength in other VQA approaches [14], [15], [16]. These works are usually trained on large-scale databases such as WaterlooIVC4K [17], LIVE-Qualcomm [18] and LIVE-QVC [19].

To further improve the performance of VQA, it is imperative to establish a larger database and a more effective quality model. In this work, we observe an exponential attenuation relationship between the coding parameters of compressed videos and their subjective quality scores. It helps us to develop a large database that reduces the workload of manual
annotation. Experiments on randomly selected samples prove the high accuracy of the proposed database. Based on this database, we construct a perceptual data-driven NR-VQA model to predict compressed video quality, which is highly related to subjective score. The major contributions of our work are summarized as follows:

1) A large-scale quality database of compressed videos is developed with a novel semi-automatic subjective labeling method, which greatly reduces the workload of manual labeling.

2) A no-reference compressed video quality assessment with a 3D CNN for SpatioTemporal Feature Extraction and Evaluation (STFEE) is proposed. It results in an end-to-end model that jointly learns perceptually spatiotemporal features of compressed videos and a quality predictor.

3) Superior performance of our method is achieved against state-of-the-art quality prediction. In addition, our algorithm achieves reasonable performance in cross-database verification, which shows that our algorithm has good generalization and robustness.

The rest of this paper is organized as follows. In Section II, the related work is introduced. In Section III, we build the proposed large-scale database. In Section IV, we describe the proposed deep network for compressed video quality assessment. In Section V, the relevant experimental results are given. Finally, this paper is concluded in Section VI.

II. RELATED WORK

To evaluate the perceptual quality of compressed videos, numerous VQA metrics have been proposed, which may be divided into conventional metrics and deep metrics. In addition, the performance of deep metrics extremely depends on the quality and quantity of training databases.

A. Databases

The quality evaluation of compressed videos should accurately reflect the subjective perception of human eyes. Therefore, a compressed video quality database with subjective labeling is an indispensable factor in the construction of a compressed video quality evaluation algorithm. In general, a compressed video database contains videos with different contents and each content video producing different quality levels. It is crucial to provide a reliable subjective quality score for each video to guide the construction and training of the algorithm. Currently, there are relatively few quality evaluation databases containing compressed videos, mainly including: LIVE Video Quality Database [20], LIVE Mobile VQA Database [21], CSIQ [22], IVP [23] and WaterlooIVC4K Video Quality Database, as shown in Table I. However, they are limited in their sample sizes. Among them, WaterlooIVC4K is the largest database, which contains 1200 videos of only 20 different video contents. Therefore, it is imperative to establish a larger compressed video quality database, in which the biggest challenge is the enormous workload of human labeling.

B. Conventional Metrics

According to the availability of original reference videos, objective VQA models can be divided into three categories: Full-Reference VQA (FR-VQA), Reduce-Reference VQA (RR-VQA) and No-Reference VQA (NR-VQA). For FR-VQA, PSNR and SSIM are the most commonly-used FR-VQAs. Li et al. [24] constructed a video quality database and proposed an full-reference screen content video quality measure. Choi et al. [25] studied the relationship between temporal visual masking and subjective perception quality for local flicker in compressed videos. Korhonen et al. [26] observed the effect of packet loss and subjective perception in decoded videos. However, the above FR-VQAs are all difficult to obtain reference videos with perfect quality in practical applications. Therefore, full-reference models are generally inapplicable. In [27], the proposed RR-VQA is able to more accurately predict the quality of compressed videos by exploiting the statistical regularities of both natural videos and distorted videos. In [28], an FR-VQA was proposed based on spatiotemporal visual sensitivity. RR-VQAs also need some features of reference videos to participate in predicting video quality, and are also inapplicable. Thus, NR-VQAs are preferred. Mittal et al. [29] standardized the frame difference and fitted the product of the four directions calculated by the standardized coefficient with asymmetric generalized Gaussian distribution. Finally, shape parameters were utilized to regress the perceived quality fraction of videos. In [30], a supervised learning approach was proposed to address the NR-VQA problem. Although the advantage of NR-VQAs without any reference information is of great application value [31], [32], most of the existing NR-VQAs have the limitation of artificially extracting features, and are difficult to obtain ideal generalization performance.

C. Deep Metrics

CNNs have shown its advantage in compressed video quality assessment with promising successes in recent years, which can jointly learn features and make quality predictors. Liu et al. [1] established a large number of compressed videos to train a deep learning network model. However, the Mean Opinion Score (MOS) of the database was obtained by full reference SSIMplus [10], which is difficult to guarantee the high correlation between the MOS and objective perception. To solve the lack of data, some scholars utilized transfer learning to evaluate video quality. Li et al. [14] extracted features from the pre-trained ResNet50 to obtain perception characteristics. Chen et al. [15] adopted VGG-16 network to learn the frame-level features of videos, and then obtained the Gaussian distributed features through adversarial learning. The above algorithms achieve the extraction of spatial domain features by means of migration learning, lacking the extraction of temporal domain features.

To address this issue, it is necessary to construct a large compressed video quality evaluation database and propose a data-driven method to learn their spatiotemporal features. In this paper, a large compressed video quality database is established and a semi-automatic labeling method is adopted.
TABLE I
EXISTING COMPRESSED VIDEO DATABASES

| Databases      | Reference videos | Codecs                      | Compressed videos | Resolutions          |
|----------------|------------------|-----------------------------|-------------------|----------------------|
| LIVE Video     | 10               | MPEG-2, H.264              | 80                | 768x432              |
| LIVE Mobile VQA| 10               | H.264, HEVC                | 40                | 1280x720             |
| CSIQ           | 12               | MPEG, H.264, HEVC          | 108               | 832x480              |
| IVP            | 10               | H.264, MPEG2, Dirac Coding | 50                | 1920x1080            |
| WaterlooIVC4K  | 20               | AVC, VP9, AV1, AVS2, HEVC  | 800               | 960x540, 1920x1080   |

Fig. 1. Videos generated by the same compression method with different Qp values.

to obtain the perceptual quality score of compressed videos quickly. On this basis, STFEE algorithm is proposed.

III. PROPOSED LARGE-SCALE COMPRESSION VIDEO QUALITY DATABASE

It is still time-consuming to build a large-scale database with subjective test. Moreover, the labelling consistency and reliability should also be guaranteed while the fatigue effects should be avoided. To this aim, we develop a semi-automatic labeling method for a massive number of compressed videos. In Zhao et al. [33], a functional relationship was observed by iteratively downsample and super-resolution operations. Inspired by this, we explore the quality degradation of compressed videos and propose a quality variation law to reduce the workload of subjective labelling.

A. The Quality Variation Law in Compressed Videos

In the video lossy compression, typically, the Quantization parameter (Qp) defines the quantization step size for transform coefficients, which has a great influence on compression efficiency and video quality. In addition, the rate-distortion efficiency scheme plays a key role in video coding, which aims to achieve a tradeoff between compression efficiency and video quality distortion. Its goal is to achieve the lowest possible distortion (e.g. low Mean-Square Error (MSE) or high SSIM) at any given bit rate. In this work, we study the effect of the most important compression parameter Qp on the visual quality. Fig. 1 shows the same video frame image compressed by Versatile Video Coding (VVC) [34] with different Qp values. The software version of VVC is VTM12.0. Random Access (RA) coding structure is employed. With the increase of Qp value, the loss of details becomes more serious, for example, the roof and background become smoother, thus resulting in lower perceived quality.

To verify the relationship between Qp value and video quality, we select different scene videos encoded with different encoders (i.e. VVC, H.265/High Efficiency Video Coding (H.265/HEVC) [35], the third generation of Audio Video Coding Standard (AVS3) [36] and Hierarchical Learned Video Compression (HLVC) [37]) and different Qp values (i.e., 22, 27, 32, 37, 42, 47 and 51). The used reference software versions of VVC, HEVC and AVS3 are VTM12.0, HM9.0 and HPM6.0, respectively. RA coding structure is employed. Experimental results suggest that the MOS would decrease with the increase of Qp (higher Qp, lower quality), as shown in Fig. 2(a). Qp value is actually the serial number of quantization step ($Q_{\text{step}}$) in video coding. In order to further verify the quality variable law of compressed videos, we also observed the relationship between MOS and $Q_{\text{step}}$, as shown in Fig. 2(b). The relationship between MOS and $Q_{\text{step}}$ approximately follows an exponential decay, as shown in Eq. (1).

$$\text{MOS} = e^{-\alpha Q_{\text{step}}}, \quad (1)$$

where $Q_{\text{step}}$ is the quantization step of a compressed video, $\alpha$ is a parameter to be estimated. We normalize the quality of uncompressed videos to 1. Given a parameter $\alpha$, the quality scores of compressed videos can be quickly obtained, thereby reducing the cost of database construction.

Based on the exponential decay relationship, the workload of our subjective test is greatly reduced. A subset of videos can be labeled with reduced workload. In addition, the quality of the remaining videos can be inferred. To examine the feasibility of this semi-automatic rating, a series of related experiments are conducted as follows. Firstly, we randomly select 14 videos of different scenes. Each video is compressed with VVC and Qp {22, 27, 32, 37, 42, 47, 51}, resulting in a total of 98 compressed videos. The used reference software version of VVC is VTM12.0. Secondly, we obtain correlation evaluation of semi-automatic rating and full subjective test,
respectively. In the semi-automatic rating, 21 subjects are asked to rate a set of compressed videos. We obtain the MOS values of the remaining videos by Eq.(1). Details of this method are described in the following Section III. D. In the full subjective test, all subjects are asked to score all test videos to obtain the MOS values. Thirdly, we compare the correlation of the MOS values obtained by the above two methods.

In most VQA tasks, the MOS values tend to converge as the number of subjects increases, which is called data saturation [38]. Fig.3 shows that data saturation occurs in 19-21 subjects, where the correlation between MOS values is close to 1. Therefore, we collected effective scores of 20 groups through subjective experiments to ensure sufficient number of subjects.

The correlation evaluation of semi-automatic rating and full subjective test is presented in Table II, where Pearson Linear Correlation Coefficient (PLCC), Spearman Rank-order Correlation Coefficient (SRCC) and Kendall Rank Correlation Coefficient (KRC) are utilized as performance indicators. In addition, Fig.4 reveals the fitting curves of quantization step size and MOS, where the scatter points represent the MOS values, and the curves are the experimental results fitted according to Eq.(1). The above experimental results indicate a higher correlation between the two approaches. Clearly, the semi-automatic rating is feasible for generating our large-scale video compression quality database with the advantages of low complexity and high accuracy.

In fact, some scholars have studied the quality change law of compressed videos. The Q-STAR [39] algorithm built a video quality prediction model based on the relationship between video quality and quantization step size, temporal resolution and spatial resolution, respectively. The three parameters mean that each group of videos requires three videos to participate in subjective labeling, which is less cost-effective. In this paper, we mainly consider the relationship between quantization step and video quality. The functional relationship is
where $s$ represents quantization step size, $s_{\text{min}}$ is the selected minimum quantization step size. $Q$ denotes the quality of a video, and $\alpha$ is the parameter to be estimated.

In addition, Ma et al. [40] explored the relationship between quantization step size and frame rate to predict the quality of compressed videos. The functional relationship between the video quality and the quantization step size is:

$$Q = \frac{1 - e^{-\alpha(s_{\text{min}})}}{1 - e^{-\alpha}},$$

Equation (2)

where $c$ is the parameter to be estimated.

It is worth noting that Eqs.(1)–(3) are only used to reveal the relationship between the quality of video content and encoding parameters, which can only assist in building a database and cannot be directly applied in video quality evaluation tasks. In each encoder, for each video content, one of its MOS value and quantization step size must be known to determine the corresponding power exponential decay function, which can be done by subjective experiments in building the database. Video quality evaluation requires a fixed model for different video content and is not applicable. Therefore, a no-reference compressed video quality assessment method is desired to design for compressed videos. In order to determine the function building the database, the above three functions are adopted to predict the quality of videos, respectively. We randomly selected 14 videos of different scenes for verification experiments. Each video utilizes 7 different Qps, and a total of 98 compressed videos are obtained. The experimental procedure is the same as that of Eq.(1). To verify the feasibility of semi-automatic labeling, we randomly select a scene video with a compression level as the benchmark to obtain the parameters to be determined by Eqs.(1)–(3), respectively. Thus, three quality scores are predicted in a semi-automatic labeling manner. Finally, the above three functions are compared, and the correlation coefficient between the semi-automatic labeling result and the MOS value is shown in Table III.

| Models | PLCC   | SRCC   | KRCC   |
|--------|--------|--------|--------|
| Equation 1 | 0.9815 | 0.9619 | 0.8842 |
| Equation 2 | 0.9819 | 0.9549 | 0.8412 |
| Equation 3 | 0.9800 | 0.9515 | 0.8257 |

Experimental results indicate that the above three functions have little correlation difference between the predicted quality scores and MOS. The performance of Eqs.(2) and (3) is lower than that of the original literatures. The reason is that only the quantization step size is considered, and other parameters are ignored in order to reduce the cost of database construction. Although the three expressions are similar, there are differences in the specific expressions. Among them, Eq.(1) has the best performance, and the expression is also the simplest, so we adopt Eq.(1) to develop compressed video quality database with semi-automatic ratings.

B. Testing Video Sequences

The source videos are selected to cover diversified scenes including animals, buildings, humans, sports, plants, and landscapes. In total, 130 videos of $1920 \times 1080$ and $1280 \times 720$ resolutions are selected from Youtube and VideoSet database [41].

To further examine the representative of the video sequences, we also calculate their Spatial Information (SI) and Temporal Information (TI) values. The SI and TI were defined in ITU-T P.910 [42] to depict the maximal spatial gradient intensity and maximal temporal discontinuity of video contents, respectively. As shown in Fig.5, the maximum value of TI is close to 100, which has relatively violent motion information; the maximum value of SI exceeds 200, indicating that the video has very rich detailed information. Therefore, the selected sequences cover a vast region of SI and TI values, which are sufficiently representative and meet the requirements of the database construction.

C. Video Compression Method

The above video sequences are 8 bit and with the resolutions of $1920 \times 1080$ and $1280 \times 720$. The frame rates are 24, 25, 30, 50 and 60. The encoding process adopts VVC, AVS3 and HLVc encoders and their corresponding configuration files. In this work, the Qp values for VVC encoding are 32, 37, 42 and 47. According to the quantization steps corresponding to the Qp, the Qp values chosen for AVS3 compression are roughly the same as those for VVC, i.e., Qp=$\{39, 45, 51, 57\}$. The quality level of HLVc is mainly controlled by the hyperparameter lambda, and the four lambdas $=\{256, 512, 1024, 2048\}$ are utilized.

In addition, the used reference software versions of VVC and AVS3 are VTM12.0 and HPM6.0, respectively. RA coding structure is employed. Besides, the remaining parameters retain their default values. Encoding utilizes official codes. In total, there are $130 \times 3 \times 4 = 1560$ outputs with different contents, resolutions, and/or Qps.

D. Semi-Automatic Labeling

According to ITU-R BT.500 [43] Recommendation, settings of display equipment are adjusted according to the daily
viewing habits of subjects. The test is executed in a laboratory environment with a normal indoor lighting level. During the process, lighting remains constant and the environment keeps silent. The video player adopted is potplayer [44], which can support up to 4K resolution and turn off the filter. The host computer is configured as AMD R9 3950X. In addition, Sony KD-75Z8H LCD screen is utilized as the display device.

Our testing procedure follows the ITU-R BT.500 document with two phases. In the pre-training phase, all subjects are told about our testing procedures and trained to score videos of different quality levels. In the formal-testing phase, all subjects are asked to watch and rate 390 videos with a workload of 1.5 hours. The test sequences are presented in random order, which are displayed on a 4K screen. 23 subjects participated in the subjective experiment, including 12 males and 11 females aged between 20 to 25 with regular visual acuities. The testing procedure follows the Double Stimulus Impairment Scale (DSIS) method defined in ITU-R BT.500, where all videos are randomly sorted and presented with unimpaired references. The user grading follows the rule of Absolute Category Rating (ACR)-5 scores. The playback time of both a reference video and a test video is set to 5 seconds. The time interval is set to a 1-second black screen to remind the subjects that the reference video is finished playing and ready to play the corresponding test video. After the test video playback is completed, a 3-second scoring time is set.

After statistical analysis, the scores of 3 subjects are identified to be outliers while the remaining 20 scores are averaged to obtain the MOS values of 390 videos. Since each video corresponds to a compression level in the compressed videos generated by each encoder, its MOS value is obtained by the fully subjective test, so the parameter \(\alpha\) of the video can be obtained. With the value \(\alpha\) and \(Q_{step}\) we can derive all the inferred MOS (iMOS) values of the remaining 1170 videos by Eq.(1). The histogram of the iMOS values is shown in Fig.6, where it appears that the test compression videos well cover the full range of quality levels. By the above process, we construct the Compression Video quality database with Semi-Automatic Ratings (CVSAR), which contains a total of 1560 compression videos. Among them, only 390 videos are manually labeled with a workload of 1.5 hours per subject while the other videos are calculated by Eq.(1). By contrast, a full subjective test of all videos takes 6 hours per subject. Therefore, the semi-automatic rating significantly reduces the workload of subjective test but generates iMOS values that are highly correlated to human ratings.

### E. Summary of the CVSAR Database

Based on the above methods, we construct a so far the largest of compressed video quality database CVSAR, which covers diverse video contents as shown in Table IV, a total of 130 reference videos with resolutions of \(1280 \times 720\) and \(1920 \times 1080\). In total, there are 1560 compressed videos in the CVSAR database with different contents, resolutions, encoders and Qps. Settings of CVSAR database are presented in Table V.

### IV. PROPOSED STFEE MODEL

We propose an end-to-end no-reference quality assessment method for compressed videos. The proposed model, namely STFEE, is a 3D CNN to predict the perceptual quality of compressed videos. The network architecture is illustrated in Fig.7. Firstly, an input video sequence is equally divided into several sub-sequences of the same size, and each sub-sequence is divided into multiple small cubes of the same size in a non-overlapping manner. Then, the video sub-sequences are fed into our 3D CNN to extract spatiotemporal features, which are utilized as the input of the transformer regression network [45]. Finally, the regression network performs long-term memory-dependent learning on the spatiotemporal features of sub-sequences in different time periods and extracts the global features of the corresponding videos. The global features are regressed onto the quality score of compressed videos by fully connected layers. In the following subsections, video preprocessing, spatiotemporal feature extracting, spatiotemporal features regressing and model training will be discussed in detail.
A. Video Preprocessing

Visual saliency is an inherent attribute of Human Visual System (HVS). In addition, it is a key factor affecting video perceptual quality [46]. The advantages of introducing visual saliency into video quality assessment are primarily reflected in two aspects: first, it allocates constrained hardware resources to more significant regions; second, video quality analysis considering visual saliency is more consistent with human visual perception. Therefore, we select the improved Holistically-nested Edge Detection (HED) [47] as our video saliency model based on comprehensive comparison and analysis of popular video saliency models. The saliency model has strong applicability and its high accuracy.

Based on the saliency model, video preprocessing process is shown in Fig. 7. First, the subsequences of videos are cut every half second, and each subsequence is a continuous 16 frames. Then, salient regions are extracted by using the improved HED saliency detection algorithm. Finally, video segmentation is performed based on the minimum circumscribed rectangle. The size of video blocks is set to $224 \times 224$. Small cubes are cut along the timeline. The dimension of each cube is $224 \times 224 \times 3 \times 16$, where 3 and 16 are the number of channels and consecutive frames, respectively.

B. Spatiotemporal Feature Extraction

3D convolution kernel can effectively extract video spatiotemporal features. In this work, we select the Inflated 3D ConvNet (I3D) [48] as our video feature extraction model based on comprehensive comparison and analysis of popular 3D convolutional networks. Based on the I3D network, the spatiotemporal features of each small cube are extracted. The features of sub-sequences are obtained by pooling the spatiotemporal features of all its small cubes. Since the $5 \times 5 \times 5$ convolution kernel of the I3D network causes a large amount of computation, we utilize a $1 \times 1 \times 1$ convolution kernel for dimensionality reduction. In order to enhance the learning ability of the network, an attention mechanism is utilized to optimize feature extraction during network training. On the basis of comprehensive consideration of feature extraction and computational complexity, a channel attention module is introduced into the convolutional layers of the last two Inception Modules of the network structure.

Based on our improved I3D network, each small cube will obtain a $1 \times 1024$ dimensional feature, which is as follows:

$$F_{\text{seq}} = \text{Pool}(F_i), \ i = 1, 2, \ldots, N_{\text{cube}},$$

$$F = \text{I3D(cube)},$$

where the cube represents a small cube. $\text{I3D()}$ denotes feature extraction operation. $F$ is spatiotemporal feature of each small cube. $\text{Pool()}$ refers to the pooling operation, which pools the spatiotemporal features of all small cubes into subsequence-level spatiotemporal features, namely $F_{\text{seq}}$. $N_{\text{cube}}$ refers to the number of small cubes divided by each subsequence.

In the same subsequence, the feature vectors of several small cubes are combined into a feature map $F$ with a dimension of $N \times 1024$, where $N$ indicates the number of small cubes in the subsequence. The pooling operation is performed on the feature map $F$ to obtain spatiotemporal features of the subsequences.

Feature Pooling:

$$F_{\text{avg}} = \text{AvgPool}(F); \ \text{shape} = (1; 1024);$$

$$F_{\text{max}} = \text{MaxPool}(F); \ \text{shape} = (1; 1024);$$

$$F_{\text{seq}} = (F_{\text{avg}}; F_{\text{max}}); \ \text{shape} = (1; 2048);$$

Among them, $F_{\text{avg}}$ and $F_{\text{max}}$ are the results of performing mean pooling and maximum pooling according to the first dimension of the $F$, respectively. Then, these two feature vectors are directly concatenated into a more representative feature vector $F_{\text{seq}}$. For a 5-second video, 16 frames of video segments are intercepted every half second to obtain 10 subsequences. The $F_{\text{seq}}$ of the 10 subsequences constitutes the global spatiotemporal feature $\text{Global}_F$.

C. Model Learning

For an input compressed video $V$, the proposed STFEE network $M$ is utilized to predict the perceptual quality $Q_{\text{pred}}$ of compressed videos:

$$Q_{\text{pred}} = M(V, \theta),$$

where $\theta$ indicates all parameters of this network.

When the network is trained, each video corresponds to a subjective score (i.e., MOS). For training convenience, the label of the whole video is used as the label of each video block.
to guide the learning of the network. Denote the ground truth quality of the input as \( Q_{\text{label}} \). The training goal of the network \( M \) is to find the optimal parameter setting, so as to minimize the overall quality prediction loss between \( Q_{\text{pred}} \) and \( Q_{\text{label}} \) of all video cubes in the training dataset. We apply the MSE as loss function in the training process, which is widely utilized in various regression tasks.

\[
\text{Loss}_{\text{cube}} = \frac{1}{N} \sum_{i=1}^{N} \| Q_{\text{pred}}(i) - Q_{\text{label}}(i) \|_2, \tag{7}
\]

where \( Q_{\text{pred}}(i) \) and \( Q_{\text{label}}(i) \) refer to the predicted quality and MOS value of the \( i \)-th cube, respectively. \( N \) represents the number of input cubes. The SGD algorithm is utilized with a learning rate of 0.001.

The spatiotemporal features of each video subsequence are obtained according to Eq.(5). Further, the global spatiotemporal feature \( \text{Global}F \) of the entire video can be obtained. In order to learn its long-term dependency information, it is used as the input of the Transformer Encoder. Then, the overall quality score of the video is predicted through the fully connected layer regression. The loss function is:

\[
\text{Loss} = \frac{1}{K} \sum_{i=1}^{K} \| Q_{\text{pred}}(i) - Q_{\text{label}}(i) \|_2, \tag{8}
\]

where \( Q_{\text{pred}}(i) \) and \( Q_{\text{label}}(i) \) denote the predicted quality and ground truth of the \( i \)-th video, respectively. \( K \) is the number of input videos. The Adam algorithm is utilized with a learning rate of 0.001.

V. EXPERIMENTAL RESULTS

In this section, the performance of the proposed STFEE model is evaluated and compared with typical video quality metrics on four datasets. In addition, we analyze the effect of block sizes on the performance of spatiotemporal features. Finally, the cross-database test is also performed to further verify the generalization performance of the algorithm.

A. Experimental Setups

We train the STFEE model on the proposed CVSAR database. To verify the generality of the proposed method, we also select three publicly available databases of LIVE, CSIQ and WaterlooIVC4K for cross-database validations, as shown in Table I. During the experiments, the basic parameters and setups of all networks are kept consistent. Also, the algorithms based on deep learning have been retrained using the same dataset and epochs. The train-test ratio is 80:20. In addition, we utilize 5-fold cross-validation to evaluate the performance of STFEE. The performance shown in Table VI is the average performance across all test sets.

There is no uniform score range and type in these databases. The setups of the CVSAR are chosen as our standard in these experiments. Subjective scores of the other three databases are normalized to the range of \([0,1]\).

B. Performance Comparison

In this paper, we demonstrate the prediction performance of STFEE based on comparative experiments, ablation experiments and cross-dataset validation. During the experiments, the basic parameters of all networks, the settings of the structure, the allocation ratios of the training and test sets are the same.

1) Performance Comparison on Different Databases: To verify the performance of the proposed STFEE, it is evaluated on the CVSAR, LIVE, CSIQ and WaterlooIVC4K databases. STFEE is also compared with typical video quality metrics including PSNR, SSIM, MS-SSIM [49], 3D-PSD [9], SpEED-VQA [8], NIQE [6], VIDE0 [29], TLVQM [12], VSFA [14], MDTVSFSA [16], GSTVQA [15] and Shen’s algorithm [32] to show its performance. We obtain the source codes of these metrics from the author’s public websites. The STFEE method is trained and tested in the databases of CVSAR, LIVE, CSIQ and WaterlooIVC4K (the train-test ratio is 80:20). The other learning-based algorithms compared, TLVQM, VSFA, MDTVSFSA, GSTVQA and Shen’s algorithm are all retrained. In addition, we utilize the PLCC and SRCC as the performance indicators. The results summarized in Table VI, where the best and the second-best results are shown in bold and underlined,

### Table VI

**Performance Comparison of STFEE Methods**

| Methods      | CVSAR | LIVE | CSIQ | WaterlooIVC4K |
|--------------|-------|------|------|---------------|
|              | PLCC  | SRCC | PLCC | SRCC          | PLCC | SRCC | PLCC | SRCC |
| PSNR         | 0.2964| 0.3471| 0.5122| 0.4790 | 0.5442| 0.5651| 0.3049| 0.3097|
| SSIM         | 0.4606| 0.5673| 0.5405| 0.5863 | 0.5518| 0.6002| 0.4290| 0.4022|
| MS-SSIM      | 0.4734| 0.6991| 0.5912| 0.6485 | 0.6119| 0.7347| 0.5938| 0.5316|
| SpEED-VQA    | 0.3630| 0.6525| 0.6040| 0.7681 | 0.7372| 0.7513| 0.5328| 0.4739|
| NIQE         | 0.4358| 0.4205| 0.2917| 0.1178 | 0.4428| 0.4282| 0.1182| 0.2467|
| VIDE0        | 0.2617| 0.1829| 0.6380| 0.6057 | 0.3087| 0.1234| 0.0094| 0.0107|
| 3D-PSD       | 0.7008| 0.7042| 0.1737| 0.1633 | 0.6321| 0.5525| 0.5922| 0.5106|
| TLVQM        | 0.6923| 0.6845| 0.3795| 0.3673 | 0.6329| 0.6178| 0.7386| 0.7627|
| VSFA         | 0.7746| 0.7687| 0.4126| 0.6102 | 0.6752| 0.7324| 0.4963| 0.5056|
| MDTVSFSA     | 0.7791| 0.7619| 0.4136| 0.5404 | 0.5519| 0.5820| 0.8927| 0.8927|
| GSTVQA       | 0.7795| 0.7444| 0.7150| 0.7237 | 0.5725| 0.5806| 0.4567| 0.4391|
| Shen’s algo. | 0.7813| 0.7932| 0.7182| 0.6463 | 0.6790| 0.7369| 0.4101| 0.3496|
| STFEE        | 0.9203| 0.9098| 0.7352| 0.7451 | 0.7883| 0.7682| 0.8813| 0.8775|

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.
respectively. Experimental results have indicated the superior performance of our proposed STFEE. From Table VI, the following observations can be drawn:

Firstly, PSNR and SSIM are essentially image quality evaluations, and do not consider video motion characteristics in temporal domain. Moreover, these two indicators simply calculate the difference between image pixels and structures, without considering human perceptual characteristics. Therefore, they perform poorly in video quality evaluation. MS-SSIM is an improved SSIM method, but its performance is still not good enough.

Secondly, STFEE obtains the highest PLCC and SRCC on the CVSAR database. VSFA, MDTVSFA, GSTVQA and Shen’s algorithm [32] are also based on deep learning, which show that they have better performance than manual feature extraction.

Thirdly, compared to the CVSAR database, the LIVE database has only one resolution of 768 × 432. The proposed STFEE is still competitive, where the PLCC still remains the highest. As a natural scene statistics method, NIQE is still not suitable for the LIVE database. Neither 3DPSD nor TLVQM indicates ideal performance on the LIVE database.

Fourthly, STFEE still exhibits the best performance on the CSIQ database. As a reduced-reference quality evaluation algorithm, SpEED-VQA has excellent performance and is more suitable for the LIVE and CSIQ databases. VSFA and Shen’s algorithm [32] also have good generalization performance on the CSIQ database.

Finally, WaterlooIVC4K has more videos and higher resolution than the LIVE and CSIQ databases. Except TLVQM and MDTVSFA, the performance of other algorithms decreased. The proposed STFEE still exhibits superior performance.

To further verify the generalization performance of STFEE with other deep learning algorithms more intuitively, we plot the PLCC values of the VSFA, MDTVSFA, GSTVQA and STFEE algorithms on the four databases into a radar chart, as shown in Fig.8.

In Fig.8, each radar chart corresponds to an algorithm, and there are 4 coordinate axes, each of which represents a database. Fig.8(a) shows that the VSFA algorithm performs poorly on the LIVE and WaterlooIVC4K databases. Fig.8(b) illustrates that the MDTVSFA algorithm performs well on the WaterlooIVC4K database, but performs poorly on the LIVE and CSIQ databases. Fig.8(c) implies that GSTVQA does not perform well on the WaterlooIVC4K database. Fig.8(d) proves that the proposed STFEE achieves superior performance on all four databases and has good generalization performance. In addition, Fig.8 also indicates that the large-scale database CVSAR can provide stable data support for deep learning methods, and the four algorithms all have good performance on the CVSAR database. Although the WaterlooIVC4K database also contains thousands of videos, it only contains 20 different scenarios. The CVSAR database includes 130 video contents and has good generalization performance.

2) Video Block Size Selection: In order to verify the effect of video block size on video spatiotemporal feature extraction, we perform the following experiments, as shown in Table VII.

Experimental results indicate that when the size of video cubes is 224 × 224, its network performance is the best. The resolution of 256 × 256 reduces the learning ability of the network, and requires higher computing power of hardwares. Therefore, we choose 224 × 224 as the size of small cubes, which enables the network to learn more effective spatiotemporal features.

3) Validation Across Datasets: To further verify the generalization performance of the proposed STFEE, cross-dataset validation is performed. We utilize LIVE, CSIQ, WaterlooIVC4K and our proposed CVSAR database as training sets to obtain network models. Then, the other three datasets are used as test sets to test their performance, as shown in Table VIII.

The above experimental results show that the proposed STFEE has good generalization ability on different databases. Among them, the performance differences on different databases are mainly due to different database scales and different compression methods. The LIVE and CSIQ databases contain a small number of compressed videos, so the models trained on these two databases are not expressive enough. For the larger WaterlooIVC4K and CVSAR datasets, the performance of the STFEE is relatively better. The above experiments also suggest that the size of the dataset is a key factor affecting the network performance.

| Size   | PLCC  | SRCC  | KRCC  |
|--------|-------|-------|-------|
| 64×64 | 0.4285| 0.4124| 0.2823|
| 128×128| 0.6038| 0.5806| 0.4087|
| 224×224| 0.7391| 0.7322| 0.5387|
| 256×256| 0.6842| 0.6707| 0.4819|
### VI. CONCLUSION

In this work, we exploit the exponentially decaying relationship between quantization step size and compressed video quality, and propose a semi-automatic rating method that greatly reduces the labeling workload while maintaining high labeling accuracy. Utilizing this approach, we build the CVSAR database, which is currently the largest database for compressed videos and has the richest scene content. Then, we develop an end-to-end STFEE model for compressed videos, which adopts a 3D convolutional network for feature extraction and follows by a transformer for quality regression. By training on the CVSAR database, STFEE model outperforms the state-of-the-art VQA algorithms. Cross-database validation also reveals the generalization ability of our STFEE model. We have made the proposed CVSAR database publicly available to facilitate reproducible research.

### REFERENCES

[1] W. Liu, Z. Duanmu, and Z. Wang, “End-to-end blind quality assessment of compressed videos using deep neural networks,” in Proc. 26th ACM Int. Conf. Multimedia, Oct. 2018, pp. 546–554.

[2] A. C. Bovik, “Automatic prediction of perceptual image and video quality,” Proc. IEEE, vol. 101, no. 9, pp. 2008–2024, Sep. 2013.

[3] A. Horé and D. Ziou, “Image quality metrics: PSNR vs. SSIM,” in Proc. 20th Int. Conf. Pattern Recognit., Aug. 2010, pp. 2366–2369.

[4] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, “Image quality assessment: From error visibility to structural similarity,” IEEE Trans. Image Process., vol. 13, no. 4, pp. 600–612, Apr. 2004.

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

[6] A. Mittal, R. Soundararajan, and A. C. Bovik, “Making a ‘completely blind’ image quality analyzer,” IEEE Signal Process. Lett., vol. 20, no. 3, pp. 209–212, Mar. 2013.

[7] M. A. Saad, A. C. Bovik, and C. Charrier, “Blind prediction of natural video quality,” IEEE Trans. Image Process., vol. 23, no. 3, pp. 1352–1365, Mar. 2014.

[8] C. G. Bampis, P. Gupta, R. Soundararajan, and A. C. Bovik, “SpEEDE-QA: Spatial efficient entropic differencing for image and video quality,” IEEE Signal Process. Lett., vol. 24, no. 9, pp. 1333–1337, Sep. 2017.

[9] M. A. Aabed, G. Kwon, and G. AlRegib, “Power of temporallyspatially unified spectral density for perceptual video quality assessment,” in Proc. IEEE Int. Conf. Multimedia Expo (ICME), Jul. 2017, pp. 1476–1481.

[10] L. Lin, S. Yu, L. Zhou, W. Chen, T. Zhao, and Z. Wang, “PEA265: Perceptual assessment of video compression artifacts,” IEEE Trans. Circuits Syst. Video Technol., vol. 30, no. 11, pp. 3898–3910, Nov. 2020.

[11] K. Zeng, T. Zhao, A. Rehman, and Z. Wang, “Characterizing perceptual artifacts in compressed video streams,” in Proc. SPIE, Feb. 2014, pp. 173–182.

[12] J. Korhonen, “Two-level approach for no-reference consumer video quality assessment,” IEEE Trans. Image Process., vol. 28, no. 12, pp. 5923–5938, Dec. 2019.

[13] Y. Zhang, H. Liu, Y. Yang, X. Fan, S. Kwong, and C. C. J. Kuo, “Deep learning based just noticeable difference and perceptual quality prediction models for compressed video,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 3, pp. 1197–1212, Mar. 2022.

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

[15] B. Chen, L. Zhu, G. Li, F. Lu, H. Fan, and S. Wang, “Learning generalized spatio-temporal deep feature representation for no-reference video quality assessment,” IEEE Trans. Circuits Syst. Video Technol., vol. 32, no. 4, pp. 1903–1916, Apr. 2022.

[16] D. Li, T. Jiang, and M. Jiang, “Unified quality assessment of in-the-wild videos with mixed datasets training,” Int. J. Comput. Vis., vol. 129, no. 4, pp. 1238–1257, Jan. 2021.

[17] Z. Li et al., “AVC, HEVC, VP9, AVS2 or AV1?—A comparative study of state-of-the-art video encoders on 4K videos,” in Proc. 16th Int. Conf. Image Anal. Recognit. (ICIAR), 2019, pp. 27–29.

[18] D. Ghadiyaram, J. Pan, A. C. Bovik, A. K. Moorthy, P. Panda, and K.-C. Yang, “In-capture mobile video distortions: A study of subjective behavior and objective algorithms,” IEEE Trans. Circuits Syst. Video Technol., vol. 28, no. 9, pp. 2061–2077, Sep. 2018.

[19] Z. Sinno and A. C. Bovik, “Large-scale study of perceptual video quality,” IEEE Trans. Image Process., vol. 28, no. 2, pp. 612–627, Feb. 2019.

[20] K. Seshadrinathan, R. Soundararajan, A. C. Bovik, and L. K. Cormack, “Study of subjective and objective quality assessment of video,” IEEE Trans. Image Process., vol. 19, no. 6, pp. 1427–1441, Jan. 2010.

[21] A. K. Moorthy, L. K. Choi, G. De Veciana, and A. Bovik, “Mobile video quality assessment database,” in Proc. IEEE ICC Workshop Realizing Adv. Video Optimized Wireless Netw., Jun. 2012, vol. 6, no. 6, pp. 652–671.

[22] P. V. Vu and D. M. Chandler, “ViS3: An algorithm for video quality assessment via analysis of spatial and spatiotemporal slices,” J. Electron. Imag., vol. 23, no. 1, pp. 013016, Feb. 2014.

[23] T. Blu et al. (2018). IVP Video Quality Database. [Online]. Available: http://ivp.ee.cuhk.edu.hk/

[24] T. Li, X. Min, H. Zhao, G. Zhai, Y. Xu, and W. Zhang, “Subjective and objective quality assessment of compressed screen content videos,” IEEE Trans. Broadcast., vol. 67, no. 2, pp. 438–449, Jun. 2021.

[25] L. X. Choi and A. C. Bovik, “Video quality assessment accounting for temporal visual masking of local ficker,” Signal Process., Image Commun., vol. 67, pp. 182–198, Sep. 2018.

[26] J. Korhonen, “Study of the subjective visibility of packet loss artifacts in encoded video sequences,” IEEE Trans. Broadcast., vol. 64, no. 2, pp. 354–366, Jun. 2018.

[27] X. Yu, N. Birkbeck, Y. Wang, C. G. Bampis, B. Adsumilli, and A. C. Bovik, “Predicting the quality of compressed videos with pre-existing distortions,” IEEE Trans. Image Process., vol. 30, pp. 7511–7526, 2021.

[28] H. Fu, D. Pan, and P. Shi, “Full-reference video quality assessment based on spatiotemporal visual sensitivity,” in Proc. Int. Conf. Culture-Oriented Sci. Technol. (ICCST), Nov. 2021, pp. 305–309.

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

[30] S. R. V. Dendi and S. S. Channappaya, “No-reference video quality assessment using natural spatiotemporal scene statistics,” IEEE Trans. Image Process., vol. 29, pp. 5612–5624, 2020.
[31] M. A. Usman, M. R. Usman, and S. Y. Shin, “A novel no-reference metric for estimating the impact of frame freezing artifacts on perceptual quality of streamed videos,” IEEE Trans. Multimedia, vol. 20, no. 9, pp. 2344–2359, Sep. 2018.

[32] W. Shen et al., “An end-to-end no-reference video quality assessment method with hierarchical spatiotemporal feature representation,” IEEE Trans. Broadcast., vol. 68, no. 3, pp. 651–660, Apr. 2022.

[33] T. Zhao, Y. Lin, Y. Xu, W. Chen, and Z. Wang, “Learning-based quality assessment for image super-resolution,” IEEE Trans. Multimedia, vol. 24, pp. 3570–3581, 2022.

[34] J. Chen and E. Alshina, Algorithm Description for Versatile Video Coding and Test Model 2 (VTM 2), document JVET-K1002-v2, 2018.

[35] G. J. Sullivan, J.-R. Ohm, W.-J. Han, and T. Wiegand, “Overview of the high efficiency video coding (HEVC) standard,” IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 12, pp. 1649–1668, Dec. 2012.

[36] J. Zhang, C. Jia, M. Lei, S. Wang, S. Ma, and W. Gao, “Recent development of AVS video coding standard: AVS3,” in Proc. Picture Coding Symp. (PCS), Nov. 2019, pp. 1–5.

[37] R. Yang, F. Mentzer, L. Van Gool, and R. Timofte, “Learning for video compression with hierarchical quality and recurrent enhancement,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2020, pp. 6627–6636.

[38] W. Zhang and H. Liu, “Toward a reliable collection of eye-tracking data for image quality research: Challenges, solutions, and applications,” IEEE Trans. Image Process., vol. 26, no. 5, pp. 2424–2437, May 2017.

[39] Y.-F. Ou, Y. Xue, and Y. Wang, “Q-STAR: A perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions,” IEEE Trans. Image Process., vol. 23, no. 6, pp. 2473–2486, Jun. 2014.

[40] Z. Ma, M. Xu, Y.-F. Ou, and Y. Wang, “Modeling of rate and perceptual quality of compressed video as functions of frame rate and quantization stepsize and its applications,” IEEE Trans. Circuits Syst. Video Technol., vol. 22, no. 5, pp. 671–682, May 2012.

[41] H. Wang et al., “VideoSet: A large-scale compressed video quality dataset based on JND measurement,” J. Vis. Commun. Image Represent., vol. 46, pp. 292–302, Jul. 2017.

[42] Methodology for the Subjective Assessment of the Quality of Television Pictures, International Telecommunication Union, document ITU-T P.910, 1999.

[43] Methodology for the Subjective Assessment of the Quality of Television Pictures, International Telecommunication Union, document ITU-R BT.500-13, 2012.

[44] Daum Communications. Accessed: Mar. 2020. [Online]. Available: https://download.cnet.com/PotPlayer-64-bit/3000-13632_4-7618571.html

[45] A. Vaswani et al., “Attention is all you need,” in Proc. Neural Inf. Process. Syst. (NIPS), 2017, pp. 5998–6008.

[46] J. Yang, C. Ji, B. Jiang, W. Lu, and Q. Meng, “No reference quality assessment of stereo video based on saliency and sparsity,” IEEE Trans. Broadcast., vol. 64, no. 2, pp. 341–353, Jun. 2018.

[47] S. Xie and Z. Tu, “Holistically-nested edge detection,” in Proc. IEEE Int. Conf. Comput. Vis. (ICCV), Dec. 2015, pp. 1395–1403.

[48] J. Carreira and A. Zisserman, “Quo vadis, action recognition? A new dataset based on JND measurement,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jul. 2017, pp. 4724–4733.

[49] Z. Wang, E. P. Simoncelli, and A. C. Bovik, “Multiscale structural similarity for image quality assessment,” in Proc. 37th Asilomar Conf. Signals, Syst. Comput., vol. 2, Jul. 2003, pp. 1398–1402.

Liqun Lin (Member, IEEE) received the M.S. and Ph.D. degrees in communication and information system from Fuzhou University, Fuzhou, China, in 2007 and 2019, respectively. She is currently an Associate Professor at Fuzhou University. Her research interests include image/video signal processing, visual quality assessment, and video coding.

Zheng Wang received the B.S. degree in information management and information system from Chongqing Jiaotong University, Chongqing, China, in 2019. He is currently pursuing the M.S. degree in electronics and communication engineering with Fuzhou University, Fujian, China. His research interests include computer vision and image quality assessment.

Jiachen He is currently pursuing the B.S. degree in electronics and communication engineering with Fuzhou University, Fujian, China. His research interests include video signal processing and visual quality assessment.

Weiling Chen (Member, IEEE) received the B.S. and Ph.D. degrees in communication engineering from Xiamen University, Xiamen, China, in 2013 and 2018, respectively. From September 2016 to December 2016, she was a Visiting Researcher at the School of Computer Science and Engineering, Nanyang Technological University, Singapore. She is currently an Associate Professor with the College of Physics and Information Engineering, Fuzhou University, China. Her current research interests include image quality assessment, image compression, and underwater acoustic communication.

Yiwen Xu (Member, IEEE) received the Ph.D. degree from the Department of Electronic Engineering, Xiamen University, Xiamen, China, in 2012. He has been a Professor with the College of Physics and Information Engineering, Fuzhou University, Fujian, China, since 2022. His research interests include AI-based multimedia information processing, video codec and transmission, and video quality assessment.

Tiesong Zhao (Senior Member, IEEE) received the B.S. degree in electrical engineering from the University of Science and Technology of China, Hefei, China, in 2006, and the Ph.D. degree in computer science from the City University of Hong Kong, Hong Kong, in 2011. He was a Research Associate with the Department of Computer Science, City University of Hong Kong, from 2011 to 2012, a Post-Doctoral Fellow with the Department of Electrical and Computer Engineering, University of Waterloo, from 2012 to 2013, and a Research Scientist with the Ubiquitous Multimedia Laboratory, The State University of New York at Buffalo, from 2014 to 2015. He is currently a Minjiang Distinguished Professor with the College of Physics and Information Engineering, Fuzhou University, China. He also serves as the Director of the Fujian Key Laboratory for Intelligent Processing and Wireless Communication of Media Information. His research interests include multimedia communication and computer vision. Due to his contributions in multimedia coding and transmission, he received the Fujian Science and Technology Award for Young Scholars in 2017. He has been serving as an Associate Editor for IET Electronics Letters since 2019.

Authorized licensed use limited to the terms of the applicable license agreement with IEEE. Restrictions apply.