Construction of VR Video Quality Evaluation Model Based on 3D-CNN

Hongxia Zhao[1], Li Huang[2]
Information Engineering College, Jiangxi University of Technology, Nanchang, China

***Abstract*—Currently, virtual reality (VR) panoramic video content occupies a very important position in the content of virtual reality platforms. The level of video quality directly affects the experience of platform users, and there is increasing research on methods for evaluating VR video quality. Therefore, this study establishes a subjective evaluation library for VR video data and uses viewport slicing method to segment VR videos, expanding the sample size. Finally, a classification prediction network structure was constructed using a three-dimensional convolutional neural network (3D-CNN) to achieve objective evaluation of VR videos. However, during the research process, it was found that the increase in its convolutional dimension inevitably leads to a significant increase in the parameter count of the entire neural network, resulting in a surge in algorithm time complexity. In response to this defect, research and design dual 3D convolutional layers and improve 3D-CNN based on residual networks. Based on this research, a virtual reality video quality evaluation model based on improved 3D-CNN was constructed. Through experimental analysis, it can be concluded that the average overall accuracy value of the constructed model is 95.27%, the average accuracy value is 95.94%, and the average Kappa coefficient value is 96.18%. Being able to accurately and effectively evaluate the quality of virtual reality videos and promote the development of the virtual reality field.**

Keywords*—Virtual reality video; 3D convolutional neural network; residual network; quality evaluation

I. INTRODUCTION

VR panoramic video is a video shot at a full 360° angle with a panoramic camera or cameras. VR panoramic video technology is a real-scene virtual reality technology based on panoramic images, which effectively integrates computer graphics technology, computer simulation technology, sensor technology, display technology and other scientific technologies [1]. VR panoramic video technology covers a variety of content and various forms of video. Among them, the video forms favoured by modern youths, including movies and recorded short films, etc., can be completed by using VR panoramic video technology [2]. At first, VR panoramic video technology was only used for leisure and entertainment, and it was rarely used in other fields. With the continuous development of technology, VR panoramic video technology has entered the fields of education, medicine, tourism, etc. [3]. The video information in mainstream VR is spherical video information, and spherical video has higher requirements for clarity, presence, vertigo, and immersion due to user experience, so timely evaluate VR video quality and improve video design become particularly important [4]. The video quality evaluation can intuitively indicate the quality of the video. 3D convolutional neural network (3D-CNN) is based on a two-dimensional convolutional neural network, adding a time dimension to the input of the neural network, extracting time and space features at the same time, performing deep learning for behaviour recognition, recognition processing and other operations method, which is widely used in video processing [5]. The research uses the viewport-cutting method to expand the number of videos. Finally, the data set is put into the training of the video quality evaluation model based on 3D-CNN. During the research process, it is found that the increase in the convolution dimension will lead to a sharp increase in the time complexity of the algorithm. In response to this defect, the study designs a double 3D convolutional layer and improves the 3D-CNN based on the shortcut of ResNet. Therefore, a VR video quality evaluation model based on improved 3D-CNN is constructed. Realize accurate and efficient objective evaluation of VR video.

This paper is divided into three parts. The first part is literature review, which analyzes the current research situation at home and abroad in the related research fields involved in the research, summarizes the existing research deficiencies, and points out the future research directions. The second part is the research method part, which expounds the VR video quality evaluation model constructed by the research institute and improves the technology used. The third part is the performance analysis part, which carries out a series of experiments to verify the performance of the model.

The importance and innovation of the research are as follows:

Traditional VR video quality evaluation methods are mainly divided into subjective and objective evaluation. The factors that affect users' perception of VR video quality are not only the perceived quality of the video, but also the subjective feelings such as presence, vertigo and accessibility. However, at present, there are few relevant subjective evaluation databases, and the traditional subjective evaluation is time-consuming and labor-intensive. However, the objective evaluation method needs to compound the subjective evaluation scores and has strong usability. The existing objective evaluation method is more complicated in calculation and its accuracy is not ideal. Therefore, a VR video quality evaluation method based on 3D-CNN was constructed. The research has two innovations, one is to establish a subjective evaluation database of VR video, which includes the quality of perception, the sense of presence, the sense of glare and the acceptability. Two: The existing VR video quality evaluation methods are full reference or partial reference. Firstly, the 3D convolutional channel network is used to evaluate the quality of VR video. Firstly, a visual

interface cutting method is proposed, and a non-reference VR visual frequency quality evaluation method is established by combining 3D convolutional channel network. This method does not need to participate in video, is driven by pure data, does not use human features extraction, and the obtained prediction results are in high consistency with the quality evaluation of V-R video, and the prediction results are good. Compared with the existing full-reference VR video quality evaluation methods, it has stronger competitiveness.

II. Related Works

With the gradual rise of VR videos, it is becoming more and more important to make accurate judgments on the quality of VR videos. Many scholars have conducted research on video quality evaluation. Tu et al. take user-generated content (UGG) videos as the research object, and comprehensively evaluate the leading no-reference/blind VQA (BVQA) features and models on a fixed evaluation framework. By employing a feature selection strategy on the BVQAmodel, a new fusion-based quality estimator for AIDeo (VIDEVAL) [6] was created. In order to compare the 8K high-resolution image quality of Versatile Video Coding (VVC) and High Efficiency Video Coding (HEVC) standards, Bonnineau et al. used PSNR, NS-SSIM and VMAF metrics for objective measurement, and obtained the comparative quality evaluation results of the two [7]. Zhang et al. aimed at the problem that many current predictive video quality of experience (QoE) models are too dependent on features specific to a specific feature set and lack generalization capabilities. Using word embedding and 3D convolutional neural networks to extract generalized features and learn them in neural networks, a new end-to-end framework (DeepQoE) was developed to perform classification and regression problems on different multimedia data [8]. Tian et al. used radial symmetric transformations on the luminance components of reference and distorted LF images to explore the depth features of geometric information in LF images. Symmetry and depth features are compared for similarity measurement to obtain video quality scores. It proposed a new full-reference image quality assessment (IQA) method [9]. Lee et al. extracted 3D shear transformation-based spatio-temporal features from overlapping video blocks and applied them to logistic regression, connected with conditional video-based deep residual neural networks to learn spatio-temporal correlations and predict quality scores. It proposed a new frequency-free reference quality assessment method [10]. Yang Aiming at the limitation of the traditional VQA method in capturing complex global time information in a panoramic video, combined spherical convolutional neural network (CNN) and non-local neural network, proposed an end-to-end neural network model for panoramic video and Stereo panoramic video for quality assessment [11].

3D-CNN is a deep learning method based on two-dimensional convolutional neural network, adding a time dimension to the input of neural network, extracting time and space features at the same time, and performing operations such as behaviour recognition and recognition processing. Mzoughi et al. proposed an efficient and fully automatic deep multi-scale three-dimensional convolutional neural network (3D-CNN) architecture based on 3D convolutional layers and deep networks to classify glioma brain tumors [12]. Ramzan et al. proposed a one-line network for segmenting multiple brain regions based on 3D-CNN, using residual learning and dilated convolution operations to learn an end-to-end mapping from MRI volumes to voxel-level brain segments [13]. The 3D convolutional neural network designed by Liu et al. for the development of Deep-Fake detection has large parameters, resulting in serious memory and storage consumption problems. A lightweight 3D-CNN [14] for DeepFake detection is proposed by using the channel transformation module to extract parameters with fewer features and fusing the spatial features on the temporal dimension for the spatio-temporal module. Hassan-Harrirou et al. In order to reduce the time and cost of exploring the chemical search space in the development of new drugs, more quickly and accurately predict the binding affinity of the lead. An ensemble of 3D-CNNs (RosENet) was used to predict the absolute binding affinities of protein-ligand complexes [15]. Aldoj et al constructed a new semi-automatic prostate cancer classification model using 3D convolutional neural network and histological correlation analysis based on multiparameter magnetic resonance (MR) imaging [16]. Salama et al. took human emotion recognition as the research goal, used the 3D-CNN deep learning framework to extract spatiotemporal features from electroencephalogram (EEG) signals and face video data, and obtained fusion predictions using data enhancement and ensemble learning techniques. The study proposed a new framework for multimodal human emotion recognition [17].

According to the comprehensive literature, there are many video quality evaluation methods, among which 3D-CNN is widely used, but it is not used in 3D video quality evaluation. Therefore, the research builds a VR video quality evaluation model based on 3D-CNN. Evaluate video quality objectively and thoughtfully.

III. Construction of VR Video Quality Evaluation Model Based on 3D-CNN

A. Evaluation Index Selection and Video Cutting

VR video is different from traditional video that only records image information from a specific angle per frame. Panoramic video captures image information from all directions at the same time. Through a professional VR head-mounted display, the video is mapped to a spherical surface for users to observe and obtain an immersive experience [18-19]. The VR video transmission framework is shown in Fig. 1.

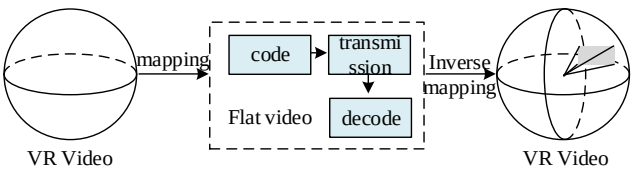


Fig. 1. VR video transmission framework.

In order to analyze video information more efficiently, the research establishes a VR video quality evaluation model based on 3D-CNN based on the characteristics of the human visual system, gives quantitative indicators to analyze video information, and obtains corresponding scores to simulate the

results of subjective evaluation and scoring. Due to the spherical 360° video of VR video, the planar (peak signal-to-noise ratio) PSNR method is usually not accurate enough, so some objective evaluation methods for VR video quality are needed. The study uses Spearman's rank correlation coefficient (SRCC) and Pearson correlation coefficient (PCC) as the correlation evaluation index of subjective and objective evaluation of video quality. The calculation formula of SRCC is shown in the following (1):

$$r_s = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n\left(n^2 - 1\right)} \tag{1}$$

In (1), d_i represents the first i difference, which n is the number of data. The PCC calculation formula is shown in the following (2).

$$\rho(x,y) = \frac{E(XY) - E(X)E(Y)}{\sqrt{E\left(X^2\right) - E^2(X)}\sqrt{E\left(Y^2\right) - E^2(Y)}} \tag{2}$$

In (2), $E(X)$ is X the mean value of the variable, and is the mean value $E(Y)$ of the Y variable. The planar video quality evaluation method is not comprehensive. In addition, the study uses spherical-based CP-PSNR and WS-PSNR to obtain more objective evaluation indicators. The weight calculation formula of ERP mapping is shown in the following (3).

$$W(i,j) = \frac{w(i,j)}{\sum_{i=0}^{W-1}\sum_{j=0}^{H-1} w(i,j)} \tag{3}$$

In (3), W and H are the length and width of the video resolution, respectively, and are the scaling factors $w(i,j)$ for pixels to (i,j) be mapped from a plane to a sphere using ERP mapping. The formula is as follows: (4).

$$w(i,j) = \cos\left(\left(j - \frac{H}{2} + \frac{1}{2}\right)\cdot\frac{\pi}{H}\right) \tag{4}$$

The calculation formula of WS-PSNR is shown in (5) below.

$$\begin{cases} WMSE = \sum_{i=0}^{W-1}\sum_{j=0}^{H-1}\left(y(i,j) - y'(i,j)^2 \cdot W(i,j)\right) \\ WS - PSNR = 10\log\left(\frac{MAX^2}{WMSE}\right) \end{cases} \tag{5}$$

In (5), $y(i,j)$ and $y'(i,j)$ are respectively the original pixel and the reconstructed pixel, which MAX is the maximum value of the color of the image point. The calculation formula of CP-PSNR is shown in the following (6).

$$\begin{cases} \bar{w}'(s,t) = \frac{w'(s,t)}{\sum_{s,t} w'(s,t)} \\ CP - PSNR = 10\log\frac{I_{\max}^2}{\sum_{s,t}\left(I(s,t) - I'(s,t)\right)^2 \cdot \bar{w}'(s,t)} \end{cases} \tag{6}$$

In (6), $I(s,t)$ and represent $I'(s,t)$ the intensity of $I_{\max}$ the point in the reference video and the damaged video respectively, which (s,t) is the maximum value of the color of the image point. In addition to the above-mentioned quality evaluation standards used in general videos, the study establishes a subjective evaluation library for VR videos, and scores the sense of presence, vertigo, and acceptability. A total of 48 VR videos were established in the research database, including 12 source reference videos. Each reference video generated 36 damaged videos through three kinds of QP, and 40 subjects were selected to participate in the establishment of the subjective evaluation database. After the subjects are trained, they evaluate the corresponding VR videos with scores. The evaluation indicators include perceptual quality, sense of presence, vertigo and acceptability. The scores given by the subjects are collected for the evaluation obtained by establishing an objective evaluation model in the future. The score (MOS) is compared with the objective evaluation of the subjects, and a corresponding scoring table is established for subsequent training. Due to the large amount of data required for deep learning, it is necessary to expand the video database, research on cutting VR videos, and cut them into small pieces of video. VR video needs to convert between spherical model and planar model. Before VR video transmission, the spherical model is mapped to the planar model. When the user watches the VR video through the HMD, the planar model is re-projected into the spherical model, so the user actually sees these are the viewports of the VR video. After VR video is projected, oversampling will occur. Based on this difference, the study uses the unique viewport characteristics of VR videos to propose a viewport cutting method to cut VR videos. The viewport segmentation diagram is shown in Fig. 2.

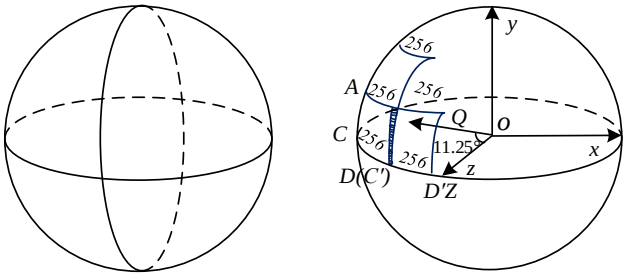


Fig. 2. Schematic diagram of viewport cutting.

Assuming that the user's viewing direction is the negative axis of the X axis and set as the initial position of the head, set the rotation matrix representing the user's rotation relative to the initial position, R and transform the three-dimensional and two-dimensional homogeneous coordinates through intrinsic matrix modeling, such as (7) as shown.

$$K=\begin{bmatrix} f_x & 0 & C_x \\ 0 & f_y & C_y \\ 0 & 0 & 1 \end{bmatrix} \tag{7}$$

In (7), C_x and C_y represent the center point of the viewport texture coordinates, f_x and f_y is the focal length expressed in pixels. The projection relationship formula is shown in the following (8).

$$w\cdot e'=K\cdot R^T\cdot E \tag{8}$$

In (8), E is a point on the spherical surface in the current visible area, e' indicating the two-dimensional homogeneous coordinates of the point mapped on the viewport, and w is the scale factor. By projecting the VR video onto a spherical surface, a series of viewports of VR videos are extracted, and the extracted viewport videos are used as the input of the VR video quality evaluation network.

B. Evaluation Model Construction based on Improved 3D-CNN

After selecting the evaluation index and establishing the evaluation library, the study uses 3D-CNN to evaluate the quality of VR video. The computing modules of traditional convolutional neural networks are mainly divided into four types: convolution, pooling, full connection and classifier prediction. The spatial structure features are extracted through the convolutional layer, and the larger-scale feature learning results are formed through the pooling layer, and then the depth of the convolutional layer and the pooling layer is continuously deepened to extract the spatial features of the image [20-21]. Convolutional neural network (CNN) was originally designed for the feature extraction of two-dimensional data, which can directly establish the mapping relationship from low-level features to high-level semantic features, and has achieved remarkable results in the field of two-dimensional image classification. However, 2D-CNN only carries out sliding calculation on the two-dimensional plane and cannot carry out feature extraction on the spectral dimension of hyperspectral image, so the extracted information is insufficient. A large number of theories and experiments show that 3D-CNN can extract features from both spatial and spectral information dimensions of hyperspectral images to improve the classification performance of the network. 3D-CNN is a convolutional network with three dimensions, namely image width, image height and image channel. The convolution kernel can move in three directions, and it can be used to better capture the temporal and spatial feature information in the video. The convolution operation formula is shown in the following (9).

$$f(x,y)*w(x,y)=\sum_{s=-a}^{a}\sum_{t=-b}^{b}w(s,t)f(x-s,y-t) \tag{9}$$

In (9), it represents $f(x,y)$ the gray value of the point whose $w(x,y)$ coordinates are in the image, which (x,y) is the convolution kernel. Sliding on the image through the weight window, the weighted sum of the pixels on the image and the weight window is used to extract the advanced features of the image. The number of input and output feature maps of the pooling layer is the same, and the calculation matrix form of the pooling layer is expressed as (10).

$$vec(y)=S(x)vec(x) \tag{10}$$

In (10), $vec(\ )$ represents the vectorization operation, and $S(x)$ the feature selector matrix of the input feature map for the pooling layer. In order to prevent the fitting phenomenon, reduce the number of parameters, and perform maximum pooling on the image, the process is shown in Fig. 3.

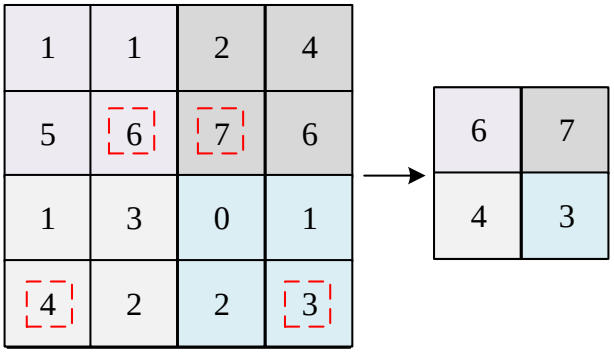


Fig. 3. Maximum pooling process.

After the pooling layer, since the output is a two-dimensional feature map, the fully connected layer vectorizes the map, and the obtained vector is as in (11).

$$a=\left[M_{1,1,1},M_{1,1,2},...,M_{1,1,n},...,M_{1,2,1},M_{1,2,2},...,M_{1,2,n},...,M_{1,m,n},M_{2,m,n},...,M_{j,m,n}\right]^T \tag{11}$$

In (11), the original feature map is $M_1,M_2,...,M_n$. The study uses the Softmax regression model at the output layer to realize the multi-category prediction function, and uses the hypothesis function to estimate the probability value of a certain class in the data set. The formula is shown in the following (12).

$$P\left(y^i=j\mid x^i;\theta\right)=\frac{e^{\theta_j^T x^i}}{\sum_{t=1}^{D}e^{\theta_j^T x^i}} \tag{12}$$

In (12), θ is the model parameter, D is the number of categories, y is the label, and x is the test input. The cost function of Softmax is as follows (13).

$$J(\theta)=-\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{k=1}^{D}1\{y^i=j\}\log\left(\frac{e^{\theta_j^T x^i}}{\sum_{t=1}^{D}e^{\theta_j^T x^i}}\right)\right]+\frac{\lambda}{2}\sum_{i=1}^{D}\sum_{j=0}^{n}\theta_{ij}^2 \tag{13}$$

In (13), $1\{\ \}$ is the indicative function, where $1\{true\}=1,1\{false\}=0$ is the coefficient of the penalty term.

λ The partial derivative of the cost function to the parameters is as follows (14).

$$\nabla_{\theta_j} J(\theta) = -\frac{1}{m}\sum_{i=1}^{m}\left[x^i 1\{y^i = j\} - P\left(y^i = j \mid x^i;\theta\right)\right] + \lambda\theta_j \tag{14}$$

The optimal solution for global convergence is obtained by minimizing the gradient descent algorithm. $J(\theta)$ Traditional convolutional neural networks can only extract spatial features in images, and are not suitable for video processing. 3D convolutional neural networks extract temporal features in videos based on traditional convolutional neural networks. And use the spatio-temporal features to classify and then predict. First of all, study the use of 3D-CNN to form a ten-category network structure. First, divide VR videos into ten categories through MOS points, of which 1-10 points are divided into one category, 10-20 points are divided into two categories, and so on, 90- 100 points are divided into 10 categories. 3D-CNN extracts temporal and spatial features based on 2D convolutional neural networks. The classification structure of 3D-CNN consists of eight 3D convolutional layers, five 3D maximum pooling layers, two fully connected layers and a ten-category output layer. After the output layer is calculated by softmax, the classification result is obtained. The research uses the cross entropy function of Softmax as the loss function, and the formula is shown in the following (15).

$$L = -\left[y\log\hat{y} + (1-y)\log(1-\hat{y})\right] \tag{15}$$

In (15), y represents the subjective MOS score obtained in the evaluation database, and $\hat{y}$ represents the predicted score. The training iterations are performed according to the loss function to obtain ten classes. After the classification is completed, use 3D-CNN to form a regression prediction model, use transfer learning, load the model parameters saved by classification as the pre-training model of the regression prediction model, and then train the regression model to give the predicted value of the video MOS score. The regression prediction structure of 3D-CNN is similar to the classification structure, the difference is that the fully connected layer of the network structure is followed by a regression prediction node. The schematic diagram of the structure is shown in Fig 4.

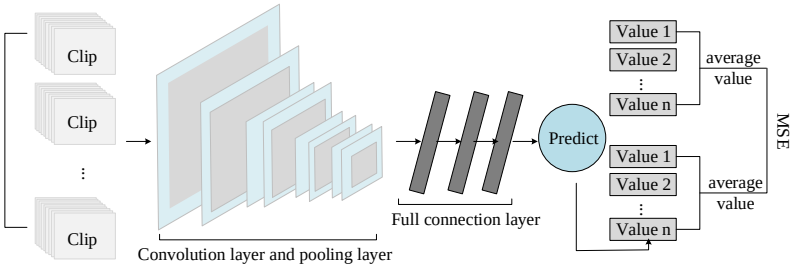


Fig. 4. Prediction structure diagram.

The parameters such as weight parameters and bias items obtained after loading the classification process through transfer learning are loaded to include all convolutional layer and pooling layer parameters, and the parameters of the two fully connected layers are discarded. The loss function uses the mean square error (MSE), and the calculation formula is as follows (16).

$$MSE = \frac{1}{N}\sum_{t=1}^{N}(y-\hat{y})^2 \tag{16}$$

In (16), N is the number of video clips, y represents the subjective MOS score obtained in the evaluation database, and $\hat{y}$ represents the predicted score. Put the predicted value into the loss function to participate in the regression training, and finally get the most suitable prediction result. Although 3D-CNN can effectively extract features from VR video data, the increase in the convolution dimension will inevitably increase the parameters of the entire neural network, resulting in a sharp increase in the time complexity of the algorithm. The conventional network input is obtained through the calculation output of the upper layer, while the residual network (ResNet) will have a "short-circuit" structure, and the data processed by the multi-layer network and the data processed by the network layer are jointly input and transmitted to in the next layer of the network. The learning goal of ResNet is to solve the residual error, as shown in (17).

$$F(x) = H(x) - x \tag{17}$$

In (17), $H(x)$ is the expected mapping of learning. The difference is obtained through (17), so that the same part before and after the cell mapping highlights the slight changes, so that the deep network structure will not degenerate while ensuring the structural performance. ResNet can reduce the complexity of parameter calculation and reduce the error caused by network depth through the unique design of shortcut. Research on improving 3D-CNN based on ResNet's shortcut. In order to reduce the nonlinear conversion operation and improve the feature extraction ability, a double 3D convolutional layer is designed, and no pooling calculation is performed between the two convolutional layers. Batch regularization calculation is performed after each convolution, and the ReLu function is used as the activation function to reduce the risk of gradient disappearance and gradient explosion. The improved 3D residual convolution unit structure is shown in Fig. 5 below:

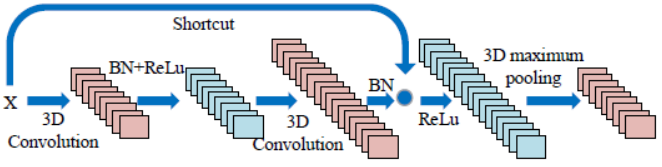


Fig. 5. Structure of 3D residual convolution unit.

Based on the above operations, the study selects video quality evaluation indicators, establishes a subjective evaluation database, uses improved 3D-CNN to classify and predict video data, obtains objective evaluation results, and uses the relationship between subjective evaluation and objective evaluation in the training process. The loss function continuously corrects and trains the model. Finally, a VR video quality evaluation model based on the improved 3D-CNN is obtained, which can accurately evaluate the VR video quality.

IV. Analysis of VR Video Quality Evaluation Model Based on 3D-CNN

An accurate and efficient video quality evaluation model is an indispensable tool to measure the pros and cons of video processing algorithms and to control video quality in real time [22]. Therefore, a VR video quality evaluation model based on improved 3D-CNN is constructed. In order to verify the performance of the constructed model, the study used three different data sets to carry out classification training on the research constructed model (model 1), the unimproved 3D-CNN model (model 2), and the BP neural network (model 3), and the convergence of the loss function as shown in Fig. 6 below:

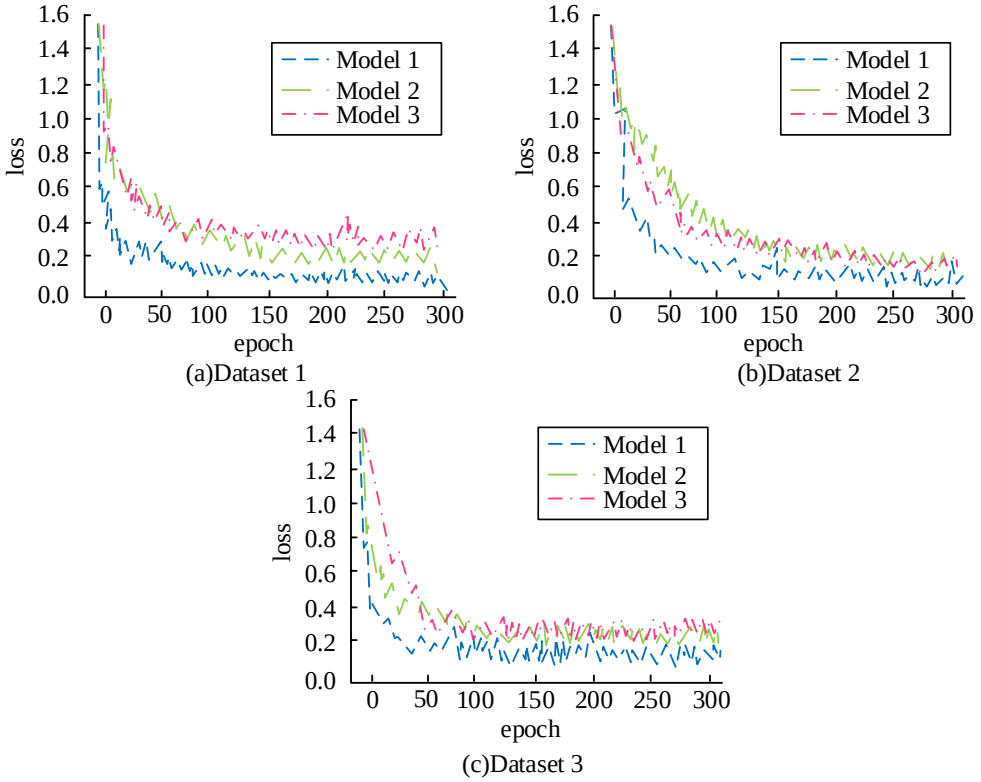


Fig. 6. Comparison of model loss function training.

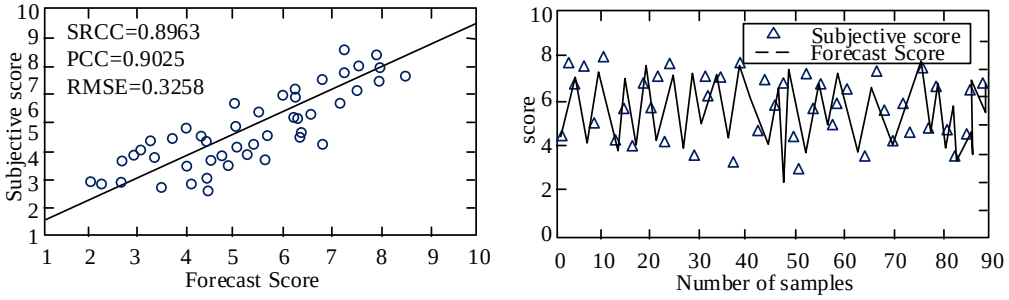
It can be seen from Fig. 6 that as the number of iterations increase, the value of the loss function of the model decreases, and gradually decreases to a certain extent. Among them, in data set 1, model 1 has the fastest decline rate and reaches the target loss value of 0.16 when the number of iterations reaches 62. However, model 2 reached the target loss value of 0.21 when iterated to 168 times, which was 0.05 higher than model 1. Model 3 reached the target loss value of 0.38 when iterated to 75 times, which was 0.21 higher than model 1; in dataset 2, model 1 reached the target loss value of 0.15 when iterated to 101 times. Model 2 reaches the target loss of 0.28 when iterating to 152 times, which is 0.13 higher than Model 1. Model 3 reaches the target loss of 0.29 after 197 iterations, which is 0.14 higher than that of model 1; in data set 3, model 1 reaches the target loss of 0.11 when iterated to 47 times, and model 2 reaches the target loss of 0.23 when iterated to 128 times. 0.12 higher than Model 1. Model 3 iterates to 149 times to reach the target loss value of 0.22, which is 0.11 higher than Model 1. Comprehensive comparison and analysis of the content in the above figure, it can be concluded that model 1 has the best convergence effect and the fastest convergence speed.

In order to further verify the improvement effect of 3D-CNN, after the model training, the test set was tested using model 1 and model 2. The results of the linear regression analysis graph obtained from the test and the predicted value and subjective score change fitting graph are shown in Fig. 7 shown.

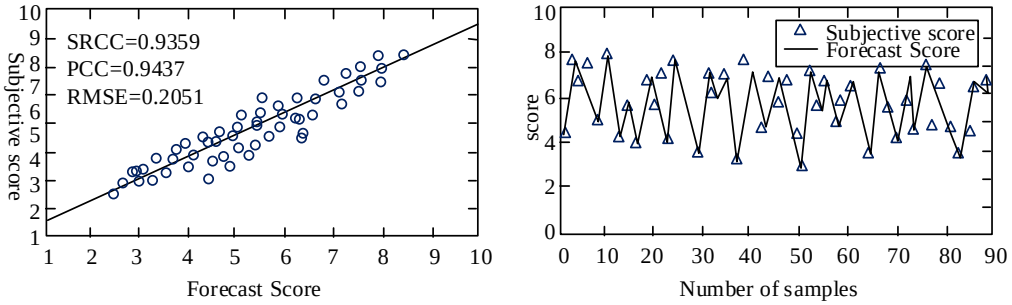
Comparing Fig. 7(a) (b), we can see that the left picture is a linear regression analysis chart, and we can see the correlation between the predicted score and the subjective score. Before the improvement, the Pearson correlation coefficient of the model was 0.9025, the Spearman coefficient

was 0.8963, and the RMSE value was 0.3258. The Pearson correlation coefficient of the improved model is 0.9437, which is 0.0412 higher than that before the improvement, the Spearman coefficient is 0.9359, which is 0.0396 higher than that before the improvement, and the RMSE value is 0.2051, which is 0.1207 lower; the right picture is the predicted value of the test sample The degree of deviation from the subjective score. It can be seen that the predicted score curve obtained by the improved model has a high degree of coincidence with the subjective score. Based on the analysis of the content in the above figure, it can be seen that the improvement of the model can effectively reduce the prediction error.

In order to compare the rationality of introducing viewport cutting into VR video clips of the model more specifically, the MOS score of each evaluation category in the subjective evaluation library before and after cutting is compared with the model prediction results, and the prediction accuracy, missed detection rate and false recognition rate are compared. The details are shown in Table I below.



(a)Before improvement



(b)After improvement

Fig. 7. Model test results before and after improvement.

TABLE I. COMPARISON AND ANALYSIS OF MODEL PREDICTION SCORE AND SUBJECTIVE SCORE

Evaluation type	Before cutting			After cutting		
	Accuracy (%)	*Undetected rate (%)*	*Error rate (%)*	*Accuracy (%)*	*Undetected rate (%)*	*Error rate (%)*
Perceived quality	84.10	11.23	15.70	93.26	9.56	7.46
Presence _	83.46	10.96	16.32	94.58	9.32	5.68
Vertigo	86.59	12.36	13.46	93.91	8.99	6.12
Acceptability	87.46	11.02	12.39	95.21	9.12	5.34
Comprehensive _	88.21	11.47	11.04	95.43	8.97	4.69

Comparative analysis of the data in Table I shows that in the process of predicting various subjective evaluation scores, the average prediction accuracy rate of the model trained before the viewport cutting method is 85.97%, the average missed detection rate is 11.41%, and the average misrecognition rate is 85.97%. The average prediction accuracy rate of the model trained after viewport cutting is 94.48%, the average missed detection rate is 9.19%, and the average false recognition rate is 5.86%. Based on the data in the above table, it can be concluded that after viewport cutting, the number of training samples is increased, and the model can evaluate video quality with higher accuracy.

In order to verify the prediction accuracy of the model and the change of running time under different sample sizes, in addition to the above three training models, the research will also commonly used classification prediction models: logistic regression (Logistics) (model 4), support vector machine (SVM) (model 5). As the number of VR video samples increases, the classification prediction accuracy of the model and the running time are shown in Fig. 8 below:

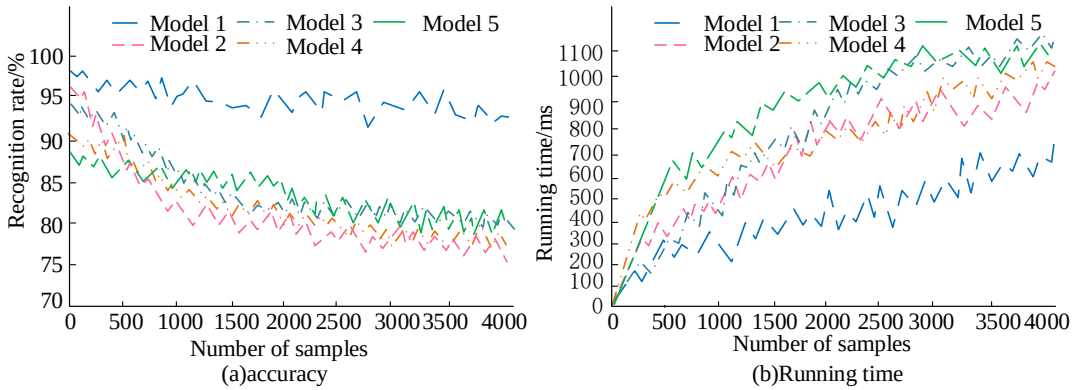


Fig. 8. Variation of prediction accuracy and running time with sample number.

It can be seen from Figure 8 that as the number of sample videos increases, the prediction accuracy of each model decreases and the running time increases. Among them, the accuracy curve of model 1 has the smallest decline. When the sample size is 500, the prediction accuracy of model 1 is 96.43%, and the running time is 0.236s. The prediction accuracy of model 2 is 90.37%, and the running time is 0.347s. The prediction accuracy of model 3 is 93.06%, and the running time is 0.343s. Model 4 has a prediction accuracy of 88.59% and a running time of 0.526s. The prediction accuracy of model 5 is 89.63%, and the running time is 0.623s; when the sample size is 4000, the prediction accuracy of model 1 is 92.64%, and the running time is 0.673s. Model 2 has a prediction accuracy of 76.85% and a running time of 0.921s. Model 3 has a prediction accuracy of 79.78% and a running time of 1.032s. The prediction accuracy of model 4 is 77.43%, and the running time is 0.996s. The prediction accuracy of model 5 is 80.12%, and the running time is 0.963s. A comprehensive analysis of the content in the above figure shows that model 1 is less affected by the increase in the number of test samples, and the running time and prediction accuracy are in a relatively ideal state.

In order to further verify the classification prediction effect of the model, the study introduces overall accuracy (OA), average accuracy (AA) and Kappa coefficient as evaluation indicators, uses five models to predict three data sets, and compares and analyzes them, as shown in Table II below.

TABLE II. EVALUATION OF MODEL CLASSIFICATION PREDICTION EFFECT

Data No.	Evaluating indicator	model 1	model 2	model 3	model 4	Model 5
Dataset 1	OA (%)	95.86	91.05	89.21	88.01	85.46
	AA (%)	97.90	92.13	88.92	87.64	86.12
	Kappa×100	95.21	91.79	89.64	86.94	85.96
Dataset 2	OA (%)	95.46	90.99	88.54	87.46	86.01
	AA (%)	94.68	91.28	89.46	88.00	85.75
	Kappa×100	96.71	92.03	88.23	87.89	86.23
Dataset 3	OA (%)	94.95	91.78	88.07	87.65	85.69
	AA (%)	95.23	90.89	89.46	87.12	86.49
	Kappa×100	96.62	91.56	88.33	88.04	85.36

Analysis of the data in Table 2 shows that the average OA value of the three data sets in model 1 is 95.27%, the average AA value is 95.94%, and the average Kappa value is 96.18%; the average OA value of the three data sets in model 2 is 90.94%, and the average AA value The value is 91.43%, and the average Kappa value is 91.79%; the average OA value of the three data sets in model 3 is 88.61%, the average AA value is 89.28%, and the average Kappa value is %; the average OA value of the three data sets in model 4 is 87.71%, the average AA value is 87.59%, the average Kappa value is 87.62%; the average OA value of the three data sets of model 5 is 85.72%, the average AA value is 86.12%, and the average Kappa value is 85.85%. Based on the data in the table, the overall accuracy (OA), average accuracy (AA) and Kappa coefficient of model 1 are higher than those of the other four models.

V. RESULTS AND DISCUSSION

With the rapid development of virtual technology and the wide application of various fields, immersive experience with a sense of presence has been widely developed. In order to evaluate the quality of VR video, the improved 3D-CNN was used to construct a VR video quality evaluation model. Through a series of experiments, the following results are obtained. Model 1 has the fastest decline speed and reaches the target loss value of 0.16 when the number of iterations reaches 62, while model 2 and model 3 both need more than 100 iterations to converge. The results show that the proposed model has good convergence performance. After the improvement, the Pearson correlation coefficient increased by 0.0412, Spearman coefficient increased by 0.0396, RMSE value decreased by 0.1207, and the predicted score curve

obtained by the improved model had a high coincidence degree with the subjective score. The results show that the improvement of the model can effectively reduce the prediction error of the model. The average prediction accuracy of the model trained after viewport cutting is 94.48%, the average missing rate is 9.19%, and the average error rate is 5.86%. This shows that the viewport cutting algorithm is reasonable and can effectively improve the model performance. The average OA value of the constructed model was 95.27%, average AA value was 95.94%, and average Kappa value was 96.18%, which could effectively and accurately evaluate the video quality. The number of VR video databases used in this study is limited, and more diverse sample videos can be found for testing and training in subsequent studies to further improve the model.

VI. Conclusion

To achieve more precise and accurate VR video quality evaluation, a VR video quality evaluation model based on improved 3D-CNN was constructed. Through the experimental analysis, it is known that the average OA value of the model constructed in the study is 95.27%, the average AA value is 95.94%, and the average Kappa value is 96.18%. It has high heterogeneity with VR video subjective quality score, and the prediction effect is better. Compared with the existing VR video quality evaluation methods, it has a strong competitiveness. In future studies, more diverse sample videos can be found for testing and training to further improve the model.

Fundings

The research is supported by Science and Technology Project of Jiangxi Provincial Department of Education: Research on VR training system of Marathon in 5G era - take the VR training system of Poyang Lake Marathon as an example (No. GJJ202009).

References

[1] M.R. Miller, F. Herrera, H. Jun, et al., "Personal identifiability of user tracking data during observation of 360-degree VR video," Scientific Reports, vol. 10, no. 1, pp. 1-10, 2020.

[2] J. Du, F.R. Yu, G. Lu, et al., "MEC-assisted immersive VR video streaming over terahertz wireless networks: A deep reinforcement learning approach," IEEE Internet of Things Journal, vol. 7, no. 10, pp. 9517 -9529, 2020.

[3] M.S. Anwar, J. Wang, W. Khan, et al., "Subjective QoE of 360-degree virtual reality videos and machine learning predictions," IEEE Access, vol. 8, pp. 148084-148099, 2020.

[4] L. Argyriou, D. Economou, V. Bouki, "Design methodology for 360 immersive video applications: the case study of a cultural heritage virtual tour," Personal and Ubiquitous Computing, vol. 24, no. 6, pp. 843-859, 2020.

[5] M. Teimouri, M. Mokhtarzade, N. Baghdadi, et al., "Fusion of time-series optical and SAR images using 3D convolutional neural networks for crop classification," Geocarto International, pp. 1-18, 2022. 2022.

[6] Z. Tu, Y. Wang, N. Birkbeck, et al., "UGC-VQA: Benchmarking blind video quality assessment for user generated content," IEEE Transactions on Image Processing, vol. 30, pp. 4449-4464, 2021.

[7] C. Bonnineau, W. Hamidouche, J. Fournier, et al., "Perceptual quality assessment of HEVC and VVC standards for 8K video," IEEE Transactions on Broadcasting, vol. 68, no. 1, pp. 246-253, 2022.

[8] H. Zhang, L. Dong, G. Gao, et al., "DeepQoE: A multimodal learning framework for video quality of experience (QoE) prediction," IEEE Transactions on Multimedia, vol. 22, no. 12, pp. 3210-3223, 2020.

[9] Y. Tian, H. Zeng, J. Hou, et al., "A light field image quality assessment model based on symmetry and depth features," IEEE Transactions on Circuits and Systems for Video Technology, vol. 31, no. 5, pp. 2046 -2050, 2020.

[10] G.Y. Lee, S.S. Shin, H. G. Kim, "No-reference sports video-quality assessment using 3D shearlet transform and deep residual neural network," Journal of Korea Multimedia Society, vol. 23, no. 12, pp. 1447-1453, 2020.

[11] J. Yang, T. Liu, B. Jiang, et al., "Panoramic video quality assessment based on non-local spherical CNN," IEEE Transactions on Multimedia, vol. 23, pp. 797-809, 2020.

[12] H. Mzoughi, I. Njeh, A. Wali, et al., "Deep multi-scale 3D convolutional neural network (CNN) for MRI gliomas brain tumor classification. Journal of Digital Imaging, vol. 33, no. 4, pp. 903- 915, 2020.

[13] F. Ramzan, M.U.G. Khan, S. Iqbal, et al., Volumetric segmentation of brain regions from MRI scans using 3D convolutional neural networks," IEEE Access, vol.8, pp. 103697-103709, 2020.

[14] J. Liu, K. Zhu, W. Lu, et al., A lightweight 3D convolutional neural network for deepfake detection," International Journal of Intelligent Systems, vol. 36, no. 9, pp. 4990-5004, 2021.

[15] H. Hassan-Harrirou, C. Zhang, T. Lemmin, "RosENet: improving binding affinity prediction by leveraging molecular mechanisms energies with an ensemble of 3D convolutional neural networks," Journal of chemical information and modeling, vol. 60, no. 6, pp. 2791-2802, 2020.

[16] N. Aldoj, S. Lukas, M. Dewey, et al., "Semi-automatic classification of prostate cancer on multi-parametric MR imaging using a multi-channel 3D convolutional neural network," European radiationology, vol. 30, no. 2, pp. 1243-1253, 2020.

[17] E.S. Salama, R.A. El-Khoribi, M.E. Shoman, et al., "A 3D-convolutional neural network framework with ensemble learning techniques for multi-modal emotion recognition," Egyptian Informatics Journal, vol. 22, no. 2, pp. 167 -176, 2021.

[18] Y. Tokuoka, T.G. Yamada, D. Mashiko, et al., "3D convolutional neural networks-based segmentation to acquire quantitative criteria of the nucleus during mouse embryogenesis," NPJ systems biology and applications, vol. 6, no. 1, pp. 1 -12, 2020.

[19] F. Fu, J. Wei, M. Zhang, et al., "Rapid vessel segmentation and reconstruction of head and neck angiograms using 3D convolutional neural network," Nature communications, vol. 11, no. 1, pp. 1 -12, 2020.

[20] J. Hong, J. Liu, "Rapid estimation of permeability from digital rock using 3D convolutional neural network," Computational Geosciences, vol. 24, no. 4, pp. 1523-1539, 2020.

[21] L. Meng, Y. Tian, S. Bu, "Liver tumor segmentation based on 3D convolutional neural network with dual scale," Journal of applied clinical medical physics, vol. 21, no. 1, pp. 144-157, 2020.

[22] D.M. Khan, N. Yahya, N. Kamel, et al., "Automated diagnosis of major depressive disorder using brain effective connectivity and 3D convolutional neural network," IEEE Access, vol. 9, no. 1, pp. 8835-8846, 2021.