Panoramic Video Quality Assessment Based on Non-local Spherical CNN

Jiachen Yang, *Member, IEEE*, Tianlin Liu, Bin Jiang, Wen Lu, *Member, IEEE*, and Qinggang Meng, *Senior Member, IEEE*,

***Abstract*—Panoramic video and stereoscopic panoramic video are essential carriers of virtual reality content, so it is very crucial to establish their quality assessment models for the standardization of virtual reality industry. However, it is very challenging to evaluate the quality of the panoramic video at present. One reason is that the spatial information of the panoramic video is warped due to the projection process, and the conventional video quality assessment (VQA) method is difficult to deal with this problem. Another reason is that the traditional VQA method is problematic to capture the complex global time information in the panoramic video. In response to the above questions, this paper presents an end-to-end neural network model to evaluate the quality of panoramic video and stereoscopic panoramic video. Compared to other panoramic video quality assessment methods, our proposed method combines spherical convolutional neural networks (CNN) and non-local neural networks, which can effectively extract complex spatiotemporal information of the panoramic video. We evaluate the method in two databases, VRQ-TJU and VR-VQA48. Experiments show the effectiveness of different modules in our method, and our method outperforms state-of-the-art other related methods.**

***Index Terms*—Virtual reality, neural network model, panoramic video, quality assessment, spatiotemporal information.**

I. Introduction

As a new means of simulation and interaction, virtual reality (VR) has attracted more and more attention in recent years [1]. Panoramic video and stereoscopic panoramic video are essential means of constructing virtual reality. Panoramic video has an unparalleled sense of realism and immersion, which can make viewers feel as if they are there. However, low quality panoramic video can cause intense discomfort and even cause physical illness [2], [3]. The process of making panoramic video and stereoscopic panoramic video is complex [4], including shooting, stitching [5], blending, projection, encoding, etc. Each process will distort the original video and affect the quality of the panoramic video [6]. Therefore, to promote the standardization of panoramic video, it is very imperative to carry out the related work of virtual reality video quality assessment (VRVQA).

This work was partially supported by National Natural Science Foundation of China (NO. 61871283), Foundation of Pre-Research on Equipment of China (NO.61403120103) and Major Civil-Military Integration Projet in Tianjin (NO.18ZXJMTG00170).(*Corresponding author: Bin Jiang.*)

B. Jiang, J. Yang and T. Liu are with School of Electrical and Information Engineering, Tianjin University, Tianjin, P.R.China (e-mail: jiangbin@tju.edu.cn; yangjiachen@tju.edu.cn; liutianlin@tju.edu.cn).

W. Lu is with School of Electronic Engineering, Xidian University, Xian, P.R.China (e-mail: luwen@xidian.edu.cn).

Q. Meng is with the Department of Computer Science, Loughborough University, Loughborough, UK (email:q.meng@lboro.ac.uk).

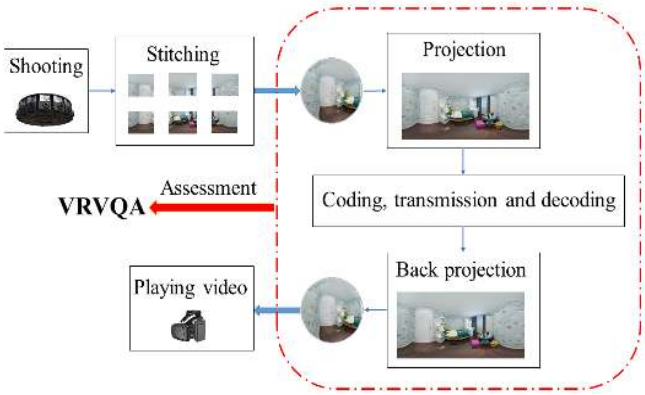


Fig. 1: The process of making panoramic video, in which VRVQA's work is mainly concentrated between the original video during viewing and the distorted video during viewing.

In the traditional multimedia quality assessment, the design process of the algorithm is often to extract features manually, and then use the machine learning method to perform regression prediction. The design of the two steps is often designed differently. In the extraction of features, the commonly used assistant theories are saliency [7]–[10] and the human visual system (HVS) [11]–[14]. Support vector regression (SVR) [15] is usually used in the regression process.

However, the panoramic video is very different from the ordinary video [16], [17]. The part of the production process and quality assessment of the panoramic video are shown in Fig. 1. When making the panoramic video, we first need to shoot with multiple panoramic cameras, and then merge the captured videos into a sphere. For ease of transmission and encoding, the panoramic video is projected from the sphere onto the plane. The panoramic video is encoded and decoded and then projected onto the sphere by a plane for viewing. Therefore, the video data we need to process is often projected to the plane, but projection will cause the original shape of regular objects to bend [18], so ordinary image quality assessment (IQA) and video quality assessment (VQA) methods are challenging to extract useful features, VRVQA must perform targeted processing on the projection process of panoramic video.

To solve the problem of projection, some methods have been proposed in VRVQA field. Xu *et al*. [19] proposed two kinds of objective assessment methods: non-content-based perceptual peak signal to noise ratio (NCP-PSNR) and content-based perceptual PSNR (CP-PSNR). The difference between the two

is whether to predict the viewing direction of the person and calculate the difference, and then perform the weight mapping of the region for the PSNR operation. Yang *et al*. [20] used 3D convolutional neural networks (CNN) to evaluate the quality of local panoramic video blocks, and then assigned different weights to combine all video blocks to obtain the overall quality of the video. Sun *et al*. [21] proposed weighted-to-spherically-Uniform PSNR (WS-PSNR), which gives all pixels different weights in advance and then calculates the PSNR. Yu *et al*. [22] projected the pixels on the original panoramic video plane and the distorted panoramic video plane onto a sphere, and then performed a large number of uniform sampling on the spherical surface to calculate the PSNR. They proposed two indicators, S-PSNR and L-PSNR, which differ in whether they give higher weight to the equator. Zakharchenko *et al*. [23] proposed the Craster parabolic projection PSNR (CPP-PSNR), which projects all the panoramic video to the sphere using the CPP method.

Although the above methods have achieved excellent results, two problems have not been effectively solved. Firstly, feature extraction in the spatial domain has to be discussed. Most of the above VRVQA methods are the improvement of traditional quality assessment methods, lacking feature extraction methods for panoramic video and advanced techniques such as deep learning. Secondly, global time domain feature extraction in the time domain has to be discussed. Global time domain information refers to the relationship between the pixels of each frame and all the pixels of other frames, which is different from finding the local time domain information of current pixels and some pixels in other frames. Most of the above VRVQA methods do not consider the effect of the relevant information of pixels in different frames of the video on the quality assessment.

Based on this, we design a deep neural network for panoramic video, which can extract the information of the panoramic video spatial domain and global time domain effectively. We verify the effectiveness of the proposed method through a series of experiments. Our main contributions are below.

1) The proposed network can effectively extract panoramic video features. Compared with ordinary CNN, spherical CNN [24] can effectively extract the "deformed" features in the panoramic video, and has translation invariance, rotation invariance and scale invariance in panoramic video processing. Spherical CNN projects the panoramic image from the plane to the three-dimensional sphere, and extracts the relevant features on the sphere by convolution. Therefore, we use spherical CNN as the basis for the proposed method. Comparative experiments verify the effectiveness of spherical CNN in VRVQA.
2) The proposed network can make full use of the global temporal information of the input data. Non-local neural networks module [25] makes the feature map in the neural network contain attention information, so the global time information of the panoramic video can be extracted together with the spherical CNN. Besides, the nonlocal neural network module uses a residual structure, which can be embedded in the spherical CNN while maintaining the same size of input and output, rather than evaluating the spatial domain and global time domain separately. Comparative experiments verify the effectiveness of non-local module in VRVQA.
3) The model we designed can evaluate not only the quality of the panoramic video, but also the quality of the stereo panoramic video. The two are only different in the preprocessing part. Experiments show that our method can provide the best quality indicator in the field of VRVQA.

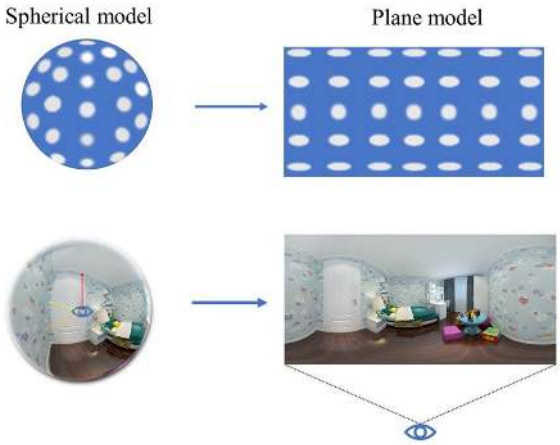


Fig. 2: The projection process of the panoramic video. In the first half of the figure, the white area represents the pixel, which changes in varying degrees after projection. The lower half of the figure is an example of projection.

In the following sections, we elaborate on the characteristics of panoramic video and related works (Section II), analyze our methods (Section III), evaluate our methods through a large number of experiments (Section IV), draw conclusions and discuss the future direction (Section V).

II. Background and Motivation

In this section, the characteristics of the space-time domain of the panoramic video and the solution ideas are introduced. Then, the description and progress of the VRVQA related work are listed separately.

A. Spatial Domain Characteristics of Panoramic Video

The projection part is the essential reason of the difference between the panoramic video and the ordinary video. To facilitate panoramic video transmission, the panoramic video must be projected onto a plane and projected back to the sphere when viewed.

There are many ways to project panoramic video [26], such as equirectangular projection (ERP) [27], cylindrical equal-area projection (EAP) [27], cube map projection (CMP) [28], rotated sphere projection (RSP) [29], etc. In respective of the method to use, it will inevitably distort the original pixel distribution and shape. ERP projection becomes the most commonly used projection method for panoramic video due to its simple processing. Similar to the projection process of

Fig. 3: Warpage deformation comparison. The three red boxes mark the same decoration in reality.

the world map, ERP stretches each latitude to the length of the equator. Due to the change of longitude of 2π and the latitude of π, the plane after the projection of the method tends to exhibit a $2:1$ aspect ratio. This method tends to stretch the two pole portions of the sphere significantly, so that the warpage deformation of the two pole portions is more severe after projection. The formula of ERP is as follows:

$$Plane(x,y) = ((\lambda - \lambda_0)cos\varphi_0, \varphi - \varphi_0)_{sphere}, \quad (1)$$

$$Sphere(\lambda,\varphi) = (\frac{x}{cos\varphi_0} + \lambda_0, y + \varphi_0)_{plane}, \quad (2)$$

where λ represents the longitude in the sphere, φ represents the latitude in the sphere, and λ_0 and φ_0 often represent the latitude and longitude of the equatorial center in the panoramic video. x and y represent the horizontal and vertical coordinates in the plane, respectively.

After projection, the original pixel distribution will be deformed, and the degree of distortion at different positions will be different. The closer to the two poles, the greater the distortion. The projection process of the panoramic video is shown in Fig. 2.

The distortion of pixels caused by projection will change with the position of the image, so the information distribution and features of the objects on the panoramic image will be quite different from the ordinary image. In general VQA method, the feature extraction operator obviously cannot adapt to the estimated deformation and extract effective features [30], and the current VRVQA field lacks the features designed for panoramic video. In order to solve this problem, a deep learning method is applied in this paper. Deep learning can automatically extract the features of panoramic video without artificial design and participation, and can extract higher-dimensional semantic information as the depth of the network increases [31]. Therefore, we design the VRVQA method based on CNN.

The same decoration in Fig. 3 has different degrees of warpage deformation at different positions after projection. This is because the displacement of an object in a spherical model belongs to three-dimensional rotation rather than translation. Therefore, the same convolution kernel is difficult to extract consistent effective features for the same object at different locations. It can be known that sparse connectivity, weight sharing, and pooling in CNN do not have translation invariance, scale invariance and rotation invariance in panoramic video [24]. For the above reasons, we must modify the convolution method in CNN, so we choose the spherical CNN that can effectively extract the features of the panoramic video.

The spherical CNN regards the spherical image as a three-dimensional manifold, expands the spherical surface into discrete three-dimensional Lie groups [32], and expresses the relationship of the special orthogonal group SO(3) in the CNN. In fact, spherical CNN can be understood as the process of convolution extraction for the input signals of three-dimensional manifold. As for the spherical CNN, this paper gives a detailed explanation in Section III, Part B. Through the above method, the two-dimensional image is reconstructed into a three-dimensional manifold, that is, the panoramic image frame is back projected back to the sphere to solve the problem of pixel deformation, and then the feature is extracted by spherical CNN. The special orthogonal group SO(3) is expressed as follows:

$$SO(3) = \left\{R \in \mathbb{R}^{3\times3} | RR^T = I, det(R) = 1\right\}, \quad (3)$$

where R represents a matrix of 3×3, and the right side of the equation indicates that the matrix is orthogonal and the determinant is 1. In the experiment, R refers to the rotation matrix, not the learnable filter parameters.

B. Global Time Domain Information Extraction of Panoramic Video

The extraction of global time domain information has always been a problem in the field of VQA. As a kind of video, panoramic video also needs to incorporate time domain information into the quality evaluation system. The previous work is divided into two categories according to whether the spatial domain information and time domain information are considered comprehensively.

The first is to extract the spatial domain and time domain information separately, and then comprehensively perform the quality assessment. Manasa *et al*. [33] used local optical flow statistics to measure the video time domain distortion to design a full reference VQA method. Zhu *et al*. [34] first obtained the characteristics of each frame of the video and then combined these features to learn the weight of the parameters from the main neural network. Ullah *et al*. [35] used long short-term memory (LSTM) to process the extracted spatial domain features, which can adequately express the information between the preceding and following frames.

The second is to extend the original 2D method to 3D, and then comprehensively consider the space-time domain information of the video. Li *et al*. [36] extended the two-dimensional discrete cosine transform (DCT) to three-dimensional, so that the space-time domain information of the video was extracted simply and effectively. Similar to the previous work, Li *et al*. [37] used the 3D shearlet transform to extract the spatiotemporal information. Giannopoulos *et al*. [38] used 3D CNN to extend the process of convolution and pooling from 2D to 3D to complete the video quality assessment.

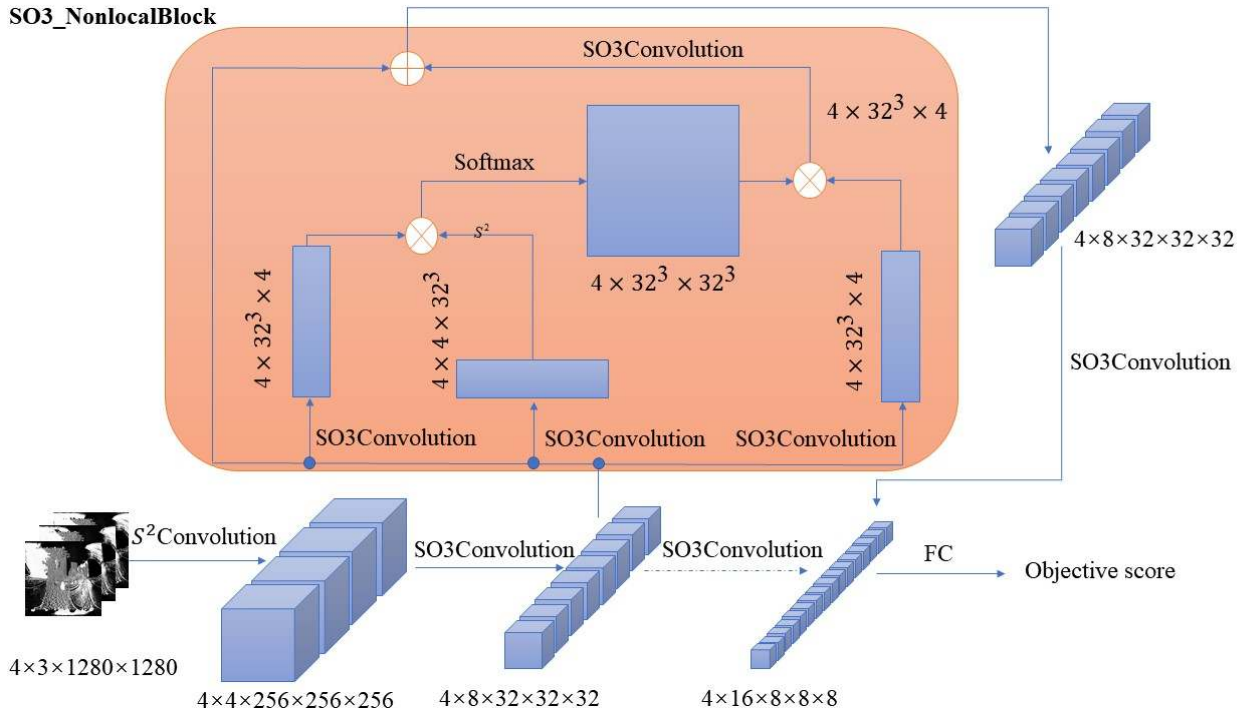


Fig. 4: The method proposed in this paper. Each convolution layer has a dimension that indicates the size of the output, and the first number "4" in the dimension represents the batch_size. The orange module describes the SO3_NonlocalBlock, and the dashed line indicates that the SO3_NonlocalBlock module is not used. "FC" indicates the complete connection layer.

The above work has brought us a lot of inspiration, but considering the cooperation with spherical CNN, an easy-to-integrate deep learning method is our best choice. CNN imitates the human cognitive process from local to macro [39]. The bottom convolution is responsible for local information and the top convolution is accountable for combining local information to get global information. However, this idea can not be applied to all situations. For example, for the quality assessment of speech video, a convolution kernel covers only around the human head. To evaluate the quality, we should not only pay attention to the distortion of the head, but also pay attention to the background of the human head, the distortion of the next frame and other related informations [40]. The same distortion appears worse on the face than on the sky, and interframe flicker distortion is also worse than continuous distortion between frames [41]. It is difficult for a single convolution layer to extract global related informations. Since the pooling process and information are transmitted layer by layer, a large amount of information is lost in the complete extraction of global information by multiple convolutional layers, so CNN has limitations in extracting global information [42].

In order to resolve the contradiction between CNN and global time domain information, non-local neural networks are integrated into our proposed framework. The non-local neural network calculates the response of a certain location as the weighted sum of the features of all positions in the input feature mapping. When we use non-local neural networks to process video, the information of each point in the feature map contains information about other points, and the input and output shapes of the non-local neural network module are the same. It is easy to insert into the neural network and can effectively extract the global temporal information of video frames.

C. General Idea of VRVQA

The quality assessment of panoramic and stereo panoramic video is in its infancy, and the related results are less than other multimedia quality evaluation fields. In order to perform VRVQA, we first need to rely on the database. Zhang *et al*. [43], Xu *et al*. [19], Zhang *et al*. [44] and Yang *et al*. [20] improved subjective assessment methods according to the characteristics of the panoramic video itself, and established a panoramic video database or a stereo panoramic video database.

Based on this work, there are two main ideas for the objective quality assessment of the panoramic video. One way is to assign different weights to different areas of panoramic video according to the pixel warping deformation or the different viewing directions of the person [19], [20]. As a representative method of this idea, WS-PSNR [21] is calculated according to the following formula:

$$WMSE = \sum_{i=0}^{w-1}\sum_{j=0}^{h-1}(y(i,j) - y'(i,j))^2 \cdot w(i,j), \quad (4)$$

$$WS-PSNR = 10log(\frac{MAX^2}{WMSE}), \quad (5)$$

$$w(i,j)_{ERP} = cos\frac{(j+0.5-h/2)\pi}{N}, \quad (6)$$

where w is the assigned weight, $y(i,j)$ and $y'(i,j)$ are the reference pixel value and the test pixel value, respectively. MAX is the maximum pixel value, h and w are the height and width of the image. N represents the total number of pixels per column.

The other way is to re-project the planar panoramic video onto the sphere and then improve the accuracy of the quality assessment by changing the projection format or sampling method [22]. As a classic method under this idea, CPP-PSNR [23] converts the video projection format into the Craster parabolic projection format to reduce the degree of pixel distortion. The formula for the CPP projection transformation is as follows:

$$Plane(x,y) = (R\lambda(2cos\frac{2\pi}{3} - 1), \pi R sin\frac{\varphi}{3})_{sphere}, \quad (7)$$

$$Sphere(\lambda,\varphi) = (\frac{x}{2cos\frac{2\varphi}{3} - 1}, \frac{3}{R} arcsin\frac{y}{\pi})_{plane}, \quad (8)$$

where φ and λ are the elevation and azimuth of the spherical coordinates, and R is the spherical radius.

The above ideas have great inspiration for our work. The method proposed in this paper mainly belongs to the second idea.

III. Proposed Method

In this section, our method is described and deduced in detail. Fig. 4 describes the primary process of our method. It should be emphasized that the method proposed in this paper can not only evaluate the quality of the stereo panoramic video, but also evaluate the quality of the ordinary panoramic video.

A. Preprocessing

Since the amount of video tends to be large, it is difficult to directly use the entire video as input to a deep learning network. The differential grayscale image can better represent stereoscopic image information based on reducing the amount of data [45], so we perform similar pretreatments. We grayscale and subtract the left and right views of the video according to the following formulas:

$$x^i = |V^i_{left} - V^i_{right}|, \quad (9)$$

where x is the output of the pre-processing, V_{left} and V_{right} represent the left and right views of the stereoscopic panoramic video frame, and i is the pixel position index.

After the above processing, the original grayscale difference map size is 2560×1280. To adapt to the input of the spherical CNN network and reduce the parameters that need to be calculated, the original grayscale difference map is downsampled to 1280×1280, and the preprocessing of spatial domain is completed.

For the preprocessing of time-domain, uniformly-spaced sampling is used. Adjacent video frames contain too much redundant information because it is difficult for the naked eye to observe the changes among them. Refer to other video research areas [46] [47], we randomly select one frame as the starting frame of the training sample video, and then extract one frame for every 8 frames. A total of 3 frames are extracted to form a video block, which is used as the input of the network. It should be noted here that due to the large amount of video data, it is difficult to read multiple consecutive frames as input during network training. The specific process is shown in Fig. 5.

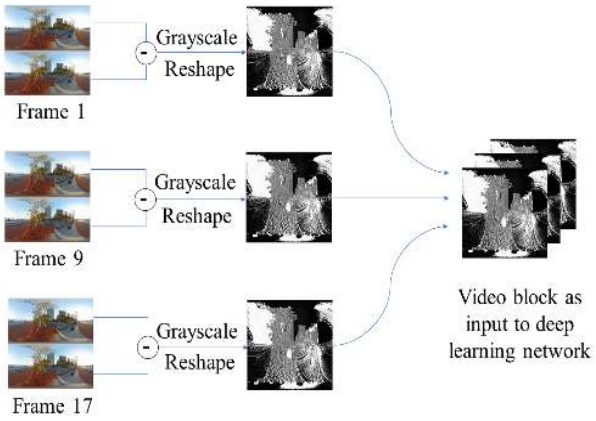


Fig. 5: The preprocessing of stereo panoramic video.

Based on the above preprocessing, a video block can be extracted every 24 frames. When processing a normal panoramic video, the input of the network becomes a grayscale image of the same size instead of a grayscale difference map, and the processing of the video block is the same.

B. Spherical CNN

In ordinary CNN, the output of the convolution operation is equivalent to the inner product of the input feature map and the convolution kernel, which is equivalent to the correlation operations in mathematics. Similar to ordinary CNN, the convolution output in spherical CNN is equivalent to the inner product of the feature map and the rotating convolution kernel, where the feature map of the spherical CNN is treated as a signal on the special orthogonal group SO(3). The convolution process of a rotating group in a spherical CNN can be expressed as follows:

$$[\varphi * f](R) =< L_R\varphi, f >= \int_{SO(3)} \sum_{i=1}^{n} \varphi_i(R^{-1}Q) f_i(Q) dQ, \quad (10)$$

where f and φ are signals on the special rotation group $SO(3) \rightarrow R^n$. $L_R\varphi$ is a rotation operator defined as $[L_R f](Q) = f(R^{-1}Q)$ on the special rotation group SO(3). dQ is a measure of the integral and can be expressed as $d\alpha sin(\beta) d\beta d\gamma/(8\pi^2)$, α, β, γ are parameters in the ZYZ Euler parameterization.

In fact, the calculation of the rotated feature map is equivalent to the inner product between the input feature map and the rotated filter.

Similar to the convolution process of a signal on SO(3), the convolution process on the surface of the sphere is called S^2. The signal convolution can be defined as:

$$[\varphi * f](R) =< L_R\varphi, f >= \int_{S^2} \sum_{i=1}^{n} \varphi_i(R^{-1}x) f_i(x) dx. \quad (11)$$

Spherical CNN uses the generalized fourier transform (GFT) to reduce the complexity of SO(3) convolution. The formulas of transform and inverse transform are expressed as follows:

$$\hat{f}^l = \int_X f(x) U^l(x) dx, \quad (12)$$

$$f(R) = \sum_{l=0}^{b} (2l+1) \sum_{m=-l}^{l} \sum_{n=-l}^{l} f_{mn}^l D_{mn}^{\hat{l}}(R), \quad (13)$$

where X refers to all manifold signals input, such as s^2 or SO3. If we input SO3, it can be understood as all spaces in three directions (α, β, γ). U^l denotes a corresponding basis function. Function f is $X \to R$, b is the bandwidth, and D is the Wigner D-functions.

C. Non-local Neural Networks

In order to extract the global time domain information of the video and process the video long-range dependencies, the non-local neural network is embedded in the designed network. The mathematical formula for non-local operations is as follows:

$$y_i = \frac{1}{C(x)} \sum_{\forall j} f(x_i, x_j) g(x_j), \quad (14)$$

where i is one of the locations of the input feature map. In general, this position can be a time point, a space point, and a space-time point. j is the index of all other possible locations, and x is the input signal, which is usually a feature map. y is the same output feature map as the x scale, f is the pairing function that calculates the correlation between the i-th position and all other positions, g is a unary input function for the purpose of information transformation, $C(x)$ is the normalization function that keeps the overall information unchanged during the conversion process.

The above functions have many manifestations. We choose to use softmax as the f function. Since softmax contains normalization process, the calculation of C is omitted, and convolution operation is used for g. 1×1 convolution operation is equivalent to matrix multiplication [25], the convolution operation is expressed as $w{\cdot}x$, w refers to the parameters of the convolution kernel update, T represents matrix transposition, then the expression of formula (14) in this paper can be converted into:

$$y_i = softmax(x^T \cdot w_{conv1}^T \cdot w_{conv2} \cdot x)(x^T \cdot w_{conv3}^T), \quad (15)$$

In order to ensure uniform size of input and output in network and more convenient configuration in the network, the design of the residual module is utilized, as shown in the formula:

$$z_i = W_z y_i + x_i, \quad (16)$$

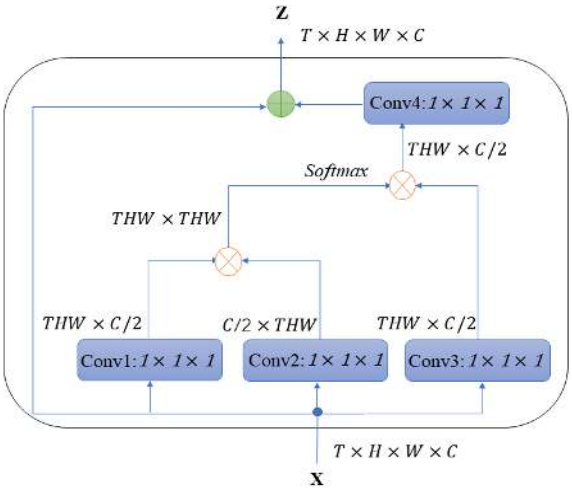


Fig. 6: Non-local neural networks example. C represents the number of channels, T represents the number of frames, and H and W represent the height and width of the feature map.

where y is the non-local operation operator and x is the input. Taking ordinary CNN as an example, the non-local operation can be as shown in the Fig. 6.

In Fig. 6, the input first passes through a convolution kernel of size $1\times1\times1$, whose main purpose is to reduce the dimension, thereby reducing the computational complexity of the non-local block. It is worth noting that this module contains the idea of attention mechanism. The softmax operation is equivalent to finding the normalized correlation between other pixels and the current pixel, and then multiplying the matrix after conv3, which applies an attention mechanism to each pixel of the input. Finally, conv4 restores the feature map to its original size and adds it to the input.

D. Network Design and Training

The epoch is set to 200. An epoch refers to the process of all data being sent to the network to perform a forward calculation and back propagation. Since an epoch is often too large and the computer can not load, we divide it into several smaller batches. Batch_Size is set to 4 because the input video size is large and limited by hardware memory. The learning rate is set to 1e-3. The number of channels in each layer is set to 4, 8, 16 and 1, and the bandwidth b is 640, 128, 32 and 8, respectively. It should be explained that the bandwidth here should be half of the input sizes H and W.

ReLU is used as an activation function in this paper. The formula is as follows:

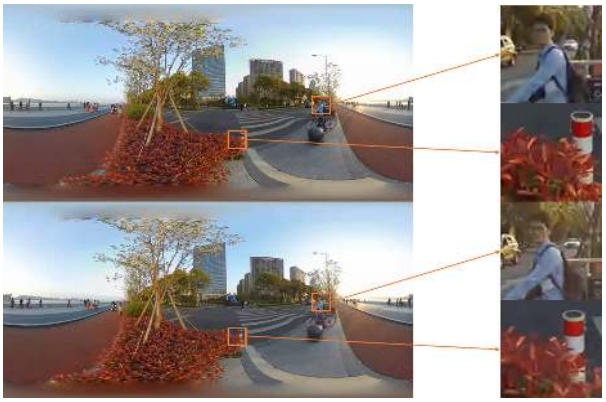
$$ReLU(x) = max(0, x), \quad (17)$$

where x refers to the input of the activation function layer.

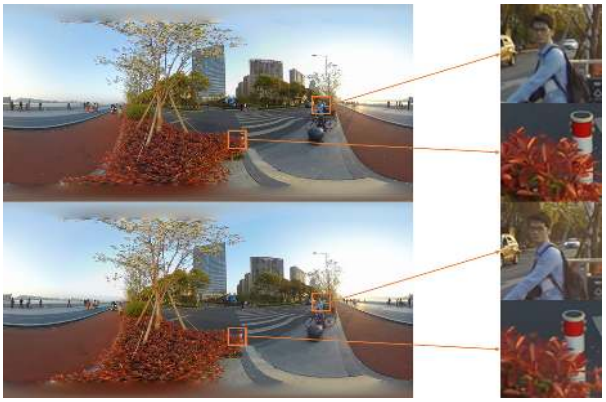
This paper chooses to use Adam [48] as the optimizer. The formula is as follows:

$$V_{dp} = [\beta_1 V_{dp-1} + (1-\beta_1) dp]/(1-\beta_1^t), \quad (18)$$

$$S_{dp} = [\beta_2 V_{dp-1} + (1-\beta_2) dp^2]/(1-\beta_2^t), \quad (19)$$



(a)



(b)

(c)

(d)

Fig. 7: (a) A sample of VRQ-TJU, which is a stereo panoramic video, MOS=1.8. (b) A sample of VRQ-TJU, which is a stereo panoramic video, MOS=4.3. (c) A sample of VR-VQA48, which is a panoramic video. MOS=38.4. (d) A sample of VR-VQA48, which is a panoramic video. MOS=62.0.

$$p = p - \alpha \frac{V_{dp}}{\sqrt{S_{dp} + \varepsilon}}, \tag{20}$$

where V_{dp} and S_{dp} are gradient first-order moment estimation and second-order moment estimation with deviation correction, respectively. α and β are attenuation factors, dp is the gradient of the parameters, and ε is the offset that prevents the denominator from being zero. In the experiment, (β_1, β_2) is (0.9, 0.99), ε is 1e-8, and α is 1e-3.

For the loss function, we choose to use the mean square error (MSE) and L2 regularization. The formula is shown below:

$$L = \frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y_i})^2 + \frac{1}{2} \lambda ||\omega||_2^2, \tag{21}$$

where L is the global loss, N is the number of all samples, it is equal to batch_size in neural network. y is the label, which is the subjective assessment score of the video. $\hat{y}$ is the predicted value, which is the objective assessment score of the video. λ is the regularization coefficient. In this paper, the regularization coefficient is equal to 0.01. w is the parameter of all layers that the network needs to update. Since the number of network layers is not large, in addition to the L2 regularization, we do not need other means to prevent over-fitting to achieve good experimental results.

In the network training phase, we use 80% of the video dataset as the training set, 20% of the video dataset as the test set, and use random sampling when segmenting data sets. In order to ensure the validity of the data, each time we repeat the experiment, we randomly divide the training set and test set again, and take the middle finger as the final result after repeating 50 experiments. When reading the training data, we randomly sample the three eligible frames in the video as input video blocks. The specific requirements are explained in the preprocessing section. Finally, we average the objective scores of 24 video blocks as the overall score of panoramic video.

IV. Experiments and Results

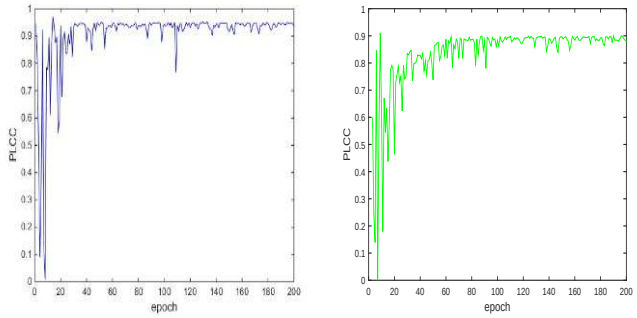
A. Datasets

Experiments are performed on VRQ-TJU[1] [20] database and VR-VQA48 [19] database to verify the effectiveness of the proposed method.

The VRQ-TJU database contains a total of 377 stereoscopic panoramic videos, including 13 original video sources. These videos are distorted by H.264 and JPEG2000. Each distortion type is divided into five levels, and each distortion type has 182 videos. Besides, the database contains symmetric distortion and asymmetric distortion, 104 of which are symmetric distortions and 260 are asymmetric distortions. Mean opinion score (MOS) is in the range [1,5], the higher the score, the better the video quality.

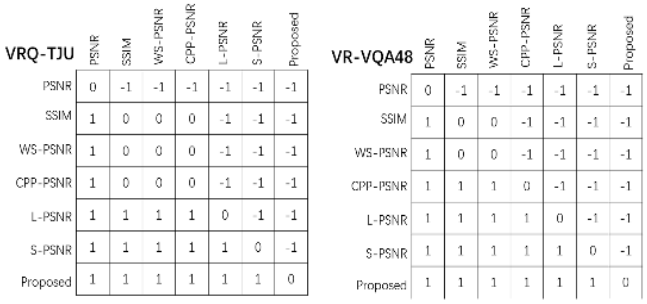
The VR-VQA48 database contains a total of 48 panoramic videos, including 12 original video sources. These videos are distorted by H.265 distortion, and the degree of distortion is divided into three levels. The MOS value is in the range of [0, 100], and the higher the score, the better the video quality.

[1]https://pan.baidu.com/s/1QDEnDARDBXDTHRcdWyPkhA



(a) (b)

Fig. 8: The relationship between the epoch and PLCC in the experiments of the two test sets. PLCC is obtained by the method proposed in this paper. (a) VRQ-TJU database. (b) VR-VQA48 database.



VRQ-TJU	PSNR	SSIM	WS-PSNR	CPP-PSNR	L-PSNR	S-PSNR	Proposed
PSNR	0	-1	-1	-1	-1	-1	-1
SSIM	1	0	0	0	-1	-1	-1
WS-PSNR	1	0	0	0	-1	-1	-1
CPP-PSNR	1	0	0	0	-1	-1	-1
L-PSNR	1	1	1	1	0	1	-1
S-PSNR	1	1	1	1	1	0	-1
Proposed	1	1	1	1	1	1	0

VR-VQA48	PSNR	SSIM	WS-PSNR	CPP-PSNR	L-PSNR	S-PSNR	Proposed
PSNR	0	-1	-1	-1	-1	-1	-1
SSIM	1	0	0	-1	-1	-1	-1
WS-PSNR	1	0	0	-1	-1	-1	-1
CPP-PSNR	1	1	1	0	-1	-1	-1
L-PSNR	1	1	1	1	0	-1	-1
S-PSNR	1	1	1	1	1	0	-1
Proposed	1	1	1	1	1	1	0

(a) (b)

Fig. 9: Results of statistical significance comparison between SROCC values from the algorithms. "1" represents the algorithm (row) better than the algorithm (column), "-1" represents the algorithm (row) worse than the algorithm (column), "0" represents similar performance.

In the VR-VQA48 database, the MOS value is often between 30 and 60.

In order to more intuitively display the video of two MOS values in two databases, we show some video frames in Fig. 7.

B. Experimental Setups

For the VRQ-TJU dataset, 302 videos are used for training and 75 videos are used for testing. For the VR-VQA48 data set, 38 videos are used for training and 10 videos are used for testing. It should be emphasized that the number of videos here is not the amount of data actually needed by the network, because 24 video blocks are proposed in each video. The comprehensive performance of the two databases can verify that the method can evaluate the quality of the panoramic video as well as the quality of the stereoscopic panoramic video. Since the data types in the two databases do not intersect with the distortion type, the two databases cannot perform cross-database experiments. Other panoramic video databases are not open source, so this paper does not involve cross-database experiments.

In the evaluation, Pearson linear correlation coefficient (PLCC) and Spearman rank order correlation coefficient (S-

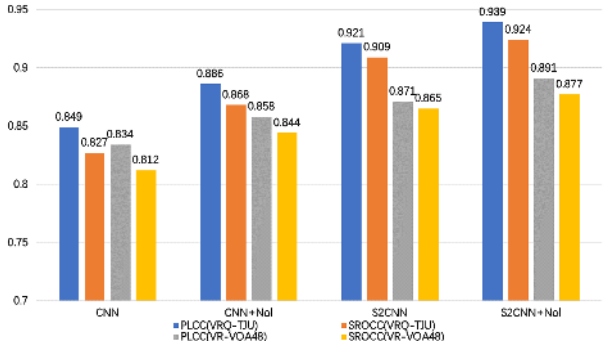


Fig. 10: Impact of different modules on results. Blue represents the PLCC index of VRQ-TJU database, grey represents the SROCC index of VRQ-TJU database, orange represents the PLCC index of VR-VQA48 database, and yellow represents the SROCC index of VR-VQA48 database. Different methods are located in abscissa.

TABLE I: Performance comparison with VRVQA methods on each database. The best performed metric is highlighted in bold type.

	VRQ-TJU		VR-VQA48	
Metrics	PLCC	SROCC	PLCC	SROCC
PSNR	0.795	0.797	0.541	0.512
SSIM [49]	0.806	0.828	0.562	0.547
WS-PSNR [21]	0.831	0.829	0.613	0.558
CPP-PSNR [23]	0.842	0.836	0.632	0.575
L-PSNR [22]	0.850	0.845	0.684	0.618
S-PSNR [22]	0.863	0.858	0.707	0.637
Proposed	**0.939**	**0.924**	**0.891**	**0.877**

ROCC) are used to predicting the accuracy. The closer the two values are to 1, the closer the objective score of the prediction is to the subjective score.

The proposed method is first tested in two databases and then compared with other classical methods. In order to verify the superiority of the spherical CNN module relative to the ordinary CNN and the effectiveness of the non-local module, relevant comparative experiments are also designed. To ensure the reliability of the experimental method and the validity of the experimental data, we repeat iterations 50 times for each analysis. In a similar experimental step, the final result tends to take the average or median of the 50 outcomes. However, the average is often affected by the outliers in the 50 data. A substantial deviation will usually give the mean has a big impact, so this paper uses the median as the final result.

C. Performance Evaluation

This section compares the proposed method with other classical VRVQA methods in two databases to prove the effectiveness of the proposed method. Our experimental environment is based on Intel(R) Xeon(R) CPU E5-2620 v4 and

TABLE III: Comparison of different modules applied in our proposed method. The best performed metric is highlighted in bold type.

	VRQ-TJU		VR-VQA48	
Metrics	PLCC	SROCC	PLCC	SROCC
CNN	0.849	0.827	0.834	0.812
CNN+Nol	0.886	0.868	0.858	0.844
S2CNN	0.921	0.909	0.871	0.865
S2CNN+Nol	**0.939**	**0.924**	**0.891**	**0.877**

NVIDIA GTX TITAN Xp GPU. The method we proposed is based on the PyTorch deep learning framework.

First, we validate the proposed method in the VRQ-TJU database and the VR-VQA48 database. In the VRQ-TJU database, the PLCC value and SROCC value of the proposed method are 0.939 and 0.924 respectively. In the VR-VQA48 database, the PLCC value and SROCC value of the proposed method are 0.891 and 0.877 respectively. Fig. 8 shows the variation of PLCC with the increase of epoch in the test set of the two databases.

In order to compare the proposed method with other VRVQA methods, we choose PSNR, SSIM [49], WS-PSNR, L-PSNR, S-PSNR, and CPP-PSNR to perform experiments in two databases. PSNR and SSIM are used as the most classic algorithms in IQA and VQA to compare with other VRVQA methods. The remaining comparison algorithms are commonly used in the field of VRVQA. The results of the experiment are shown in Table I. Among them, L-PSNR, S-PSNR and CPP-PSNR are implemented in C++, and other comparison methods are implemented in MATLAB. In Table I, we show the most advanced indicators in bold type. As can be seen from the table, our method achieves good results in both databases. In the VRQ-TJU database, our PLCC and SROCC lead 0.076 and 0.066 respectively. In the VR-VQA48 database, our PLCC and SROCC lead 0.184 and 0.24 respectively. Experiments show that our method can achieve good results in stereo panoramic video quality assessment and panoramic video quality assessment. Combining the performance of Fig. 8 and Table I, it can be found that the convergence index during training is basically consistent with the test set, indicating that the model has not been overfitted. Through experiments, the average time to train a video block in this method is 25.17 seconds, and the average time to test a video block is 0.072 seconds. Our model get the best performance with the right amount of complexity. The detailed comparison is shown in Table II.

The statistical significance of the predictions is determined by comparing the SROCC values of each VRVQA method. We assume that the predicted score follows a normal distribution, and the F-test is used to express whether the proposed method is superior to other methods. Assuming a significance level of 0.05, we calculate the result for each method using the 50 SROCC values. The value "1" indicates that the algorithm (row) is better than the algorithm (column). The value of "0" indicates statistical equivalence between rows and columns, and the value of "-1" indicates that the algorithm (row) is not as good as the algorithm (column). The results of the F-test are shown in Fig. 9. Overall, our method results are all "1", indicating that our model is superior to other models.

D. Module Comparison Evaluation

In this part, we verify the contribution of spherical CNN and non-local modules to the proposed method, and confirm the superiority of this method in the spatial domain and time domain assessment. To make the results more reliable, we conduct four experiments in two databases, using spherical CNN+ non-local modules, spherical CNN, ordinary CNN+ non-local modules, and ordinary CNN. For the sake of convenience, we write ordinary CNN as CNN, spherical CNN as S2CNN, and non-local module as Nol. In order not to add additional variables, we simply replace the corresponding structure without changing the overall settings of the hyperparameters and the network. When we want to compare spherical CNN with ordinary CNN, we only replace the corresponding layer. When we want to compare the effects of non-native modules, we only add or not add non-local modules to the network. The results of the experiment are shown in Table III, and we show the best data in bold type. In order to more intuitively show the contribution of different modules, we represent the data in the table as the form of Fig. 10.

Experiments show that both spherical CNN and non-local neural networks have significant contributions to the methods proposed in this paper, especially spherical CNN can significantly improve PLCC and SROCC. In order to fully demonstrate the correlation between objective data and subjective data obtained by different methods, we show some scatter plots in Fig. 11.

E. Distortion Type Evaluation

VRQ-TJU contains relatively many types of distortion. In order to better evaluate the performance of our method in different distortion types, we divide the two databases into five parts according to the distortion type, which are symmetric distortion database, asymmetric distortion database, H.264 distortion database, JPEG2000 distortion database, H.265 distortion database (ie VR-VQA48). Experiments are carried out in these five databases. The first four databases used the model trained by VRQ-TJU, and the last database used the model trained by VR-VQA48. The experimental data is shown in Table IV, and we show the most advanced indicators in bold type.

Experiments show that the proposed method can deliver good performance in different distortion types, and the symmetric distortion database has better results than other databases. In order to more intuitively observe the relationship between subjective scores and objective scores, we show some scatter plots of two scores in different distortion databases in Fig. 12.

TABLE II: Single frame consumption time of different VRVQA methods.

Methods	PSNR	SSIM [49]	WS-PSNR [21]	CPP-PSNR [23]	L-PSNR [22]	S-PSNR [22]	Proposed
Time(second)	0.002	0.071	0.002	0.858	0.116	0.114	0.072

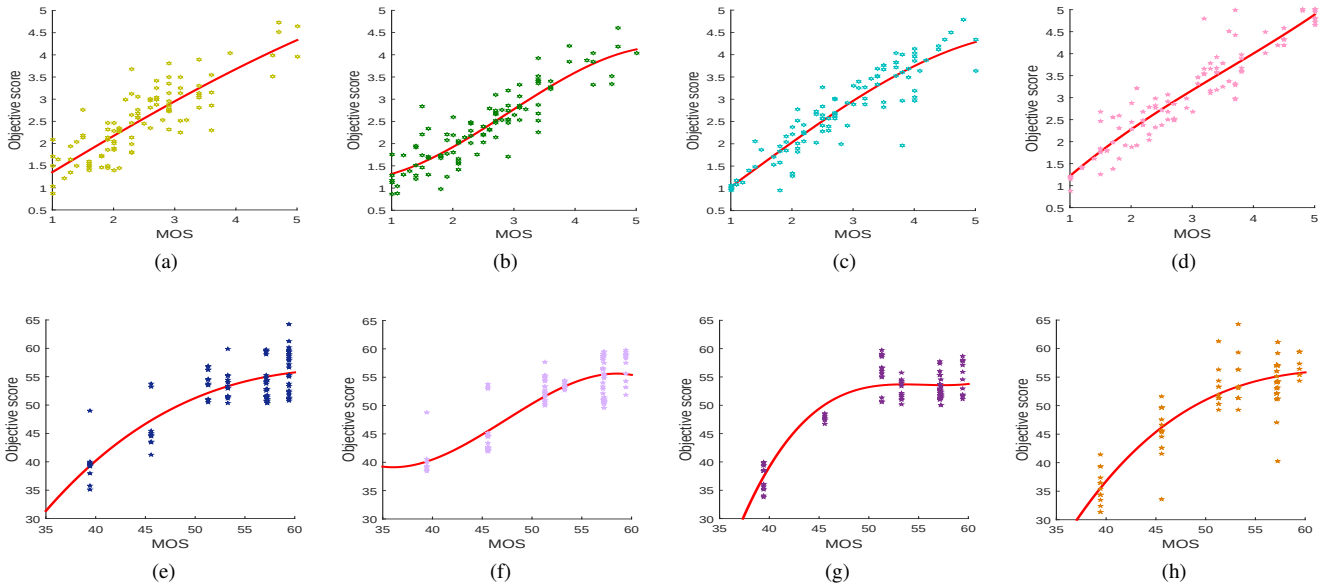


Fig. 11: The relationship between MOS and objective score based on the proposed method. (a) Using CNN in VRQ-TJU. (b) Using CNN+Nol in VRQ-TJU. (c) Using S2CNN in VRQ-TJU. (d) Using S2CNN+Nol in VRQ-TJU. (e) Using CNN on VR-VQA48. (f) Using CNN+Nolon VR-VQA48. (g) Using S2CNN on VR-VQA48. (h) Using S2CNN+Nol on VR-VQA48.

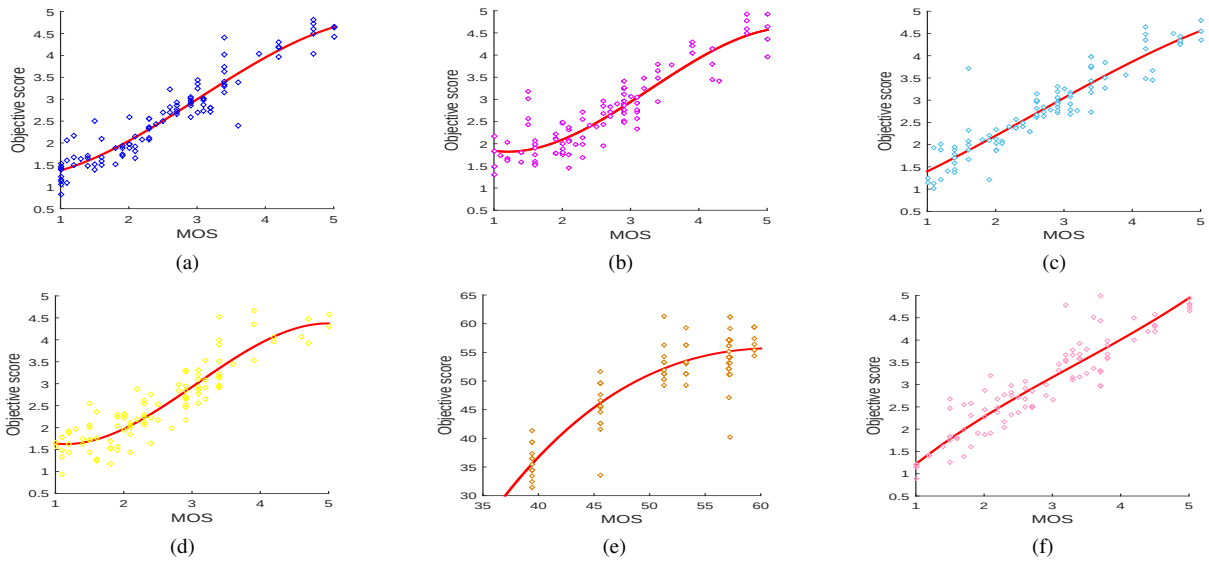


Fig. 12: The relationship between MOS and objective score based on the proposed method. (a) Symmetric distortion database. (b) Asymmetric distortion database. (c) H.264 distortion database. (d) JPEG2000 distortion database. (e) H.265 distortion database (VR-VQA48 database). (f)VRQ-TJU database.

F. Prospective Applications

(1) **Optimization of codec**

As we all know, panoramic video has the characteristics of high resolution and large amount of data, which brings considerable challenges to video coding and decoding [50]. How to measure the performance loss of codec is

TABLE IV: The performance of the proposed method and the traditional VRVQA method in database classified according to distortion type. The best performed metric is highlighted in bold type.

	Symmetric distortion		Asymmetric distortion		H.264 distortion		JPEG2000 distortion		H.265 distortion	
Metrics	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC	PLCC	SROCC
PSNR	0.845	0.857	0.772	0.767	0.781	0.775	0.724	0.715	0.541	0.512
SSIM [49]	0.866	0.883	0.764	0.805	0.746	0.763	0.726	0.802	0.562	0.547
WS-PSNR [21]	0.852	0.848	0.811	0.807	0.825	0.821	0.794	0.792	0.613	0.558
CPP-PSNR [23]	0.868	0.836	0.832	0.804	0.837	0.808	0.798	0.772	0.632	0.575
L-PSNR [22]	0.875	0.859	0.823	0.799	0.819	0.784	0.791	0.776	0.684	0.618
S-PSNR [22]	0.882	0.861	0.826	0.805	0.833	0.814	0.812	0.791	0.707	0.637
CNN	0.854	0.831	0.846	0.823	0.851	0.827	0.845	0.822	0.834	0.812
CNN+Nol	0.898	0.874	0.881	0.858	0.880	0.857	0.874	0.850	0.858	0.844
S2CNN	0.928	0.914	0.917	0.907	0.918	0.908	0.915	0.904	0.871	0.865
S2CNN+Nol	**0.946**	**0.931**	**0.930**	**0.916**	**0.937**	**0.922**	**0.929**	**0.915**	**0.891**	**0.877**

helpful to optimize the rate distortion and other related works. Therefore, in the ERP format of panoramic video coding and decoding, the method proposed in this paper is used to observe the loss of video quality, so as to provide guidance for codec.

(2) **Quality enhancement**
At present, some people have begun to study how to enhance the quality of virtual view to give viewers a better visual experience. For example, Rahaman *et al*. [51], [52] used Gaussian mixture modeling (GMM) to significantly enhance the quality of virtual views. However, in their final evaluation stage, they often do not use advanced quality evaluation methods to verify the effectiveness of the proposed methods. The method proposed in this paper can assist the verification work in the field of quality enhancement, so as to improve the accuracy of the quality enhancement method.

(3) **Standardization of hardware**
At present, the quality of virtual reality display helmets varies in the market, and non-standard hardware devices will greatly affect the user experience. The quality assessment method proposed in this paper can be used to accurately quantify the performance of hardware in video viewing, so as to unify the hardware manufacturing standards.

V. Conclusion

In this paper, we propose a method based on deep learning, which can evaluate the quality of panoramic video and stereo panoramic video end-to-end. This paper starts from the two aspects of spatial domain assessment and global time domain assessment, and studies the characteristics of panoramic video. A lot of experiments have been carried out to verify the effectiveness of the proposed method. The results show that the spherical CNN is more suitable for the extraction of panoramic video features than CNN, and the non-local neural networks module can effectively extract the global time domain information.

Although the proposed method has good results, the non-local neural networks module occupies a large number of computing resources and storage space. Due to the large amount of data in the panoramic video and the complexity of network calculations, it is difficult to use some techniques that require parameter calculation, such as the GN layer [53]. In the future work, we hope to design a more elegant time domain assessment strategy to minimize complex parameter operations based on the network performance. Cross-database experiments will also be added in the next step to verify the generalization ability of the algorithm.

References

[1] Christopher J Turner, Windo Hutabarat, John Oyekan, and Ashutosh Tiwari. Discrete event simulation and virtual reality use in industry: new opportunities and future trends. *IEEE Transactions on Human-Machine Systems*, 46(6):882–894, 2016.

[2] Laura Freina and Michela Ott. A literature review on immersive virtual reality in education: state of the art and perspectives. In *The International Scientific Conference eLearning and Software for Education*, volume 1, page 133. " Carol I" National Defence University, 2015.

[3] Ashutosh Singla, Stephan Fremerey, Werner Robitza, and Alexander Raake. Measuring and comparing qoe and simulator sickness of omnidirectional videos in different head mounted displays. In *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–6. IEEE, 2017.

[4] Shenchang Eric Chen. Quicktime vr: An image-based approach to virtual environment navigation. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 29–38. ACM, 1995.

[5] Zhe Zhu, Jiaming Lu, Minxuan Wang, Songhai Zhang, Ralph R Martin, Hantao Liu, and Shi-Min Hu. A comparative study of algorithms for realtime panoramic video blending. *IEEE Transactions on Image Processing*, 27(6):2952–2965, 2018.

[6] Andrew MacQuarrie and Anthony Steed. Cinematic virtual reality: Evaluating the effect of display type on the viewing experience for panoramic video. In *2017 IEEE Virtual Reality (VR)*, pages 45–54. IEEE, 2017.

[7] Ke Gu, Shiqi Wang, Huan Yang, Weisi Lin, Guangtao Zhai, Xiaokang Yang, and Wenjun Zhang. Saliency-guided quality assessment of screen content images. *IEEE Transactions on Multimedia*, 18(6):1098–1110, 2016.

[8] Yun Liu, Jiachen Yang, Qinggang Meng, Zhihan Lv, Zhanjie Song, and Zhiqun Gao. Stereoscopic image quality assessment method based on binocular combination saliency model. *Signal Processing*, 125:237–248, 2016.

[9] Jiachen Yang, Chunqi Ji, Bin Jiang, Wen Lu, and Qinggang Meng. No reference quality assessment of stereo video based on saliency and sparsity. *IEEE Transactions on Broadcasting*, 64(2):341–353, 2018.

[10] Wei Zhang and Hantao Liu. Study of saliency in objective video quality assessment. *IEEE Transactions on Image Processing*, 26(3):1275–1288, 2017.

[11] Jongyoo Kim, Hui Zeng, Deepti Ghadiyaram, Sanghoon Lee, Lei Zhang, and Alan C Bovik. Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment. *IEEE Signal Processing Magazine*, 34(6):130–141, 2017.

[12] Sung-Ho Bae and Munchurl Kim. A novel image quality assessment with globally and locally consilient visual quality perception. *IEEE Transactions on Image Processing*, 25(5):2392–2406, 2016.

[13] Shaoze Wang, Kai Jin, Haitong Lu, Chuming Cheng, Juan Ye, and Dahong Qian. Human visual system-based fundus image quality assessment of portable fundus camera photographs. *IEEE transactions on medical imaging*, 35(4):1046–1055, 2015.

[14] Soo-Chang Pei and Li-Heng Chen. Image quality assessment using human visual dog model fused with random forest. *IEEE Transactions on Image Processing*, 24(11):3282–3292, 2015.

[15] Bernhard Schölkopf, Alex J Smola, Robert C Williamson, and Peter L Bartlett. New support vector algorithms. *Neural computation*, 12(5):1207–1245, 2000.

[16] Alireza Zare, Alireza Aminlou, Miska M Hannuksela, and Moncef Gabbouj. Hevc-compliant tile-based streaming of panoramic video for virtual reality applications. In *Proceedings of the 24th ACM international conference on Multimedia*, pages 601–605. ACM, 2016.

[17] Christian Weissig, Oliver Schreer, Peter Eisert, and Peter Kauff. The ultimate immersive experience: panoramic 3d video acquisition. In *International Conference on Multimedia Modeling*, pages 671–681. Springer, 2012.

[18] Chi-Wing Fu, Liang Wan, Tien-Tsin Wong, and Chi-Sing Leung. The rhombic dodecahedron map: An efficient scheme for encoding panoramic video. *IEEE Transactions on Multimedia*, 11(4):634–644, 2009.

[19] Mai Xu, Chen Li, Zhenzhong Chen, Zulin Wang, and Zhenyu Guan. Assessing visual quality of omnidirectional videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[20] Jiachen Yang, Tianlin Liu, Bin Jiang, Houbing Song, and Wen Lu. 3d panoramic virtual reality video quality assessment based on 3d convolutional neural networks. *IEEE Access*, 6:38669–38682, 2018.

[21] Yule Sun, Ang Lu, and Lu Yu. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE signal processing letters*, 24(9):1408–1412, 2017.

[22] Matt Yu, Haricharan Lakshman, and Bernd Girod. A framework to evaluate omnidirectional video coding schemes. In *2015 IEEE International Symposium on Mixed and Augmented Reality*, pages 31–36. IEEE, 2015.

[23] Vladyslav Zakharchenko, Kwang Pyo Choi, and Jeong Hoon Park. Quality metric for spherical panoramic video. In *Optics and Photonics for Information Processing X*, volume 9970, page 99700C. International Society for Optics and Photonics, 2016.

[24] Taco S Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical cnns. *arXiv preprint arXiv:1801.10130*, 2018.

[25] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018.

[26] Zhenzhong Chen, Yiming Li, and Yingxue Zhang. Recent advances in omnidirectional video coding for virtual reality: Projection and evaluation. *Signal Processing*, 146:66–78, 2018.

[27] John Parr Snyder and Philip M Voxland. *An album of map projections*. Number 1453. US Government Printing Office, 1989.

[28] King-To Ng, Shing-Chow Chan, and Heung-Yeung Shum. Data compression and transmission aspects of panoramic videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(1):82–95, 2005.

[29] Adeel Abbas and David Newman. A novel projection for omni-directional video. In *Applications of Digital Image Processing XL*, volume 10396, page 103960V. International Society for Optics and Photonics, 2017.

[30] Shih-Ming Chang, Hon-Hang Chang, Shwu-Huey Yen, and Timothy K Shih. Panoramic human structure maintenance based on invariant features of video frames. *Human-Centric Computing and Information Sciences*, 3(1):14, 2013.

[31] Y Lecun, Y Bengio, and G Hinton. Deep learning. *Nature*, 521(7553):436, 2015.

[32] D. V. Alekseevskii. Lie groups. *Journal of Soviet Mathematics*, 28(6):924–949, 1985.

[33] K Manasa and Sumohana S Channappayya. An optical flow-based full reference video quality assessment algorithm. *IEEE Transactions on Image Processing*, 25(6):2480–2492, 2016.

[34] Kongfeng Zhu, Chengqing Li, Vijayan Asari, and Dietmar Saupe. No-reference video quality assessment based on artifact measurement and statistical analysis. *IEEE Transactions on Circuits and Systems for Video Technology*, 25(4):533–546, 2015.

[35] Amin Ullah, Jamil Ahmad, Khan Muhammad, Muhammad Sajjad, and Sung Wook Baik. Action recognition in video sequences using deep bi-directional lstm with cnn features. *IEEE Access*, 6:1155–1166, 2018.

[36] Xuelong Li, Qun Guo, and Xiaoqiang Lu. Spatiotemporal statistics for video quality assessment. *IEEE Transactions on Image Processing*, 25(7):3329–3342, 2016.

[37] Y. Li, L. Po, C. Cheung, X. Xu, L. Feng, F. Yuan, and K. Cheung. No-reference video quality assessment with 3d shearlet transform and convolutional neural networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 26(6):1044–1057, June 2016.

[38] Michalis Giannopoulos, Grigorios Tsagkatakis, Saverio Blasi, Farzad Toutounchi, Athanasios Mouchtaris, Panagiotis Tsakalides, Marta Mrak, and Ebroul Izquierdo. Convolutional neural networks for video quality assessment. *arXiv preprint arXiv:1809.10117*, 2018.

[39] Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

[40] Ulrich Engelke, Hagen Kaprykowsky, Hans-Jürgen Zepernick, and Patrick Ndjiki-Nya. Visual attention in quality assessment. *IEEE Signal Processing Magazine*, 28(6):50–59, 2011.

[41] Alexandre Ninassi, Olivier Le Meur, Patrick Le Callet, and Dominique Barba. Does where you gaze on an image affect your perception of quality? applying visual attention to image quality metric. In *2007 IEEE International Conference on Image Processing*, volume 2, pages II–169. IEEE, 2007.

[42] Ke Sun, Bin Xiao, Dong Liu, and Jingdong Wang. Deep high-resolution representation learning for human pose estimation. *arXiv preprint arXiv:1902.09212*, 2019.

[43] Yingxue Zhang, Yingbin Wang, Feiyang Liu, Zizheng Liu, Yiming Li, Daiqin Yang, and Zhenzhong Chen. Subjective panoramic video quality assessment database for coding applications. *IEEE Transactions on Broadcasting*, 64(2):461–473, 2018.

[44] Bo Zhang, Junzhe Zhao, Shu Yang, Yang Zhang, Jing Wang, and Zesong Fei. Subjective and objective quality assessment of panoramic videos in virtual reality environments. In *2017 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 163–168. IEEE, 2017.

[45] Wei Zhang, Chenfei Qu, Lin Ma, Jingwei Guan, and Rui Huang. Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network. *Pattern Recognition*, 59:176–187, 2016.

[46] Mohammadreza Zolfaghari, Kamaljeet Singh, and Thomas Brox. Eco: Efficient convolutional network for online video understanding. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 695–712, 2018.

[47] Limin Wang, Yuanjun Xiong, Zhe Wang, Yu Qiao, Dahua Lin, Xiaoou Tang, and Luc Van Gool. Temporal segment networks: Towards good practices for deep action recognition. In *European conference on computer vision*, pages 20–36. Springer, 2016.

[48] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[49] Zhou Wang, Alan C Bovik, Hamid R Sheikh, Eero P Simoncelli, et al. Image quality assessment: from error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.

[50] Afshin TaghaviNasrabadi, Anahita Mahzari, Joseph D Beshay, and Ravi Prakash. Adaptive 360-degree video streaming using layered video coding. In *2017 IEEE Virtual Reality (VR)*, pages 347–348. IEEE, 2017.

[51] DM Motiur Rahaman and Manoranjan Paul. Virtual view synthesis for free viewpoint video and multiview video compression using

gaussian mixture modelling. *IEEE Transactions on Image Processing*, 27(3):1190–1201, 2017.

[52] DM Motiur Rahaman, Manoranjan Paul, and Nusrat Jahan Shoumy. Virtual view quality enhancement using side view information for free viewpoint video. In *2018 IEEE Visual Communications and Image Processing (VCIP)*, pages 1–5. IEEE, 2018.

[53] Yuxin Wu and Kaiming He. Group normalization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 3–19, 2018.

Jiachen Yang (M' 15) received the M.S. and Ph.D. degrees in communication and information engineering from Tianjin University, China, in 2005 and 2009, respectively. He is currently a professor at Tianjin University. He was a visiting scholar with the department of computer science, School of Science, Loughborough University, UK. He was also a visiting scholar at Embry-Riddle Aeronautical University, Daytona Beach, FL, US. His research interests include stereo vision research, pattern recognition and image quality evaluation.

Tianlin Liu received the B.S. degree in communication and information engineering from Tianjin University, Tianjin, China, in 2017. He is currently pursuing the M.S. degree at the school of electrical and information engineering, Tianjin University, Tianjin, China. His research interests include virtual reality and multimedia quality evaluation.

Bin Jiang received the B.S. and M.S. degree in communication and information engineering from Tianjin University, Tianjin, China, in 2013 and 2016, where he is currently pursuing the Ph.D. degree. He is also a visiting scholar in Department of Electrical, Computer, Software, and Systems Engineering, Embry-Riddle Aeronautical University, Daytona Beach, FL, US, where he is a member of Security and Optimization for Networked Globe Laboratory. His research interests lie in Multimedia QoE and Security.

Wen Lu (M' 16) received the M.S. and Ph.D. degrees in electrical engineering from Xidian University, China, in 2006 and 2009, respectively. He is currently a professor at Xidian University. His research interests include image and video understanding, visual quality assessment, and computational vision.

Qinggang Meng (M' 06-SM' 18) received the B.S. and M.S. degrees from the School of Electronic Information Engineering, Tianjin University, China, and the Ph.D. degree in computer science from Aberystwyth University, UK. He is a professor with the Department of Computer Science, Loughborough University, U.K. His research interests include biologically and psychologically inspired learning algorithms and developmental robotics, service robotics, robot learning and adaptation, multi-UAV cooperation, drivers distraction detection, human motion analysis and activity recognition, activity pattern detection, pattern recognition, artificial intelligence, and computer vision. He is a Fellow of the Higher Education Academy, UK.