IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received May 27, 2018, accepted July 3, 2018, date of publication July 11, 2018, date of current version July 30, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2854922

3D Panoramic Virtual Reality Video Quality Assessment Based on 3D Convolutional Neural Networks

JIACHEN YANG[1], (Member, IEEE), TIANLIN LIU[1], BIN JIANG[1], HOUBING SONG[2], (Senior Member, IEEE), AND WEN LU[3], (Member, IEEE)

[1]School of Electrical and Information Engineering, Tianjin University, Tianjin 300072, China
[2]Department of Electrical, Computer, Software, and Systems Engineering, Embry-Riddle Aeronautical University, Daytona Beach, FL 32114, USA
[3]School of Electronic Engineering, Xidian University, Xi'an 710071, China

Corresponding author: Bin Jiang (jiangbin@tju.edu.cn)

This work was supported in part by the National Natural Science Foundation of China under Grant 61471260, in part by the Natural Science Foundation of Tianjin under Grant 16JCYBJC16000, and in part by the Foundation of Pre-Research on Equipment of China under Grant 61403120103.

ABSTRACT Virtual reality (VR), a new type of simulation and interaction technology, has aroused widespread attention and research interest. It is necessary to evaluate the VR quality and provide a standard for the rapidly developing technology. To the best of our knowledge, a few researchers have built benchmark databases and designed related algorithms, which has hindered the further development of the VR technology. In this paper, a free available data set (VRQ-TJU) for VR quality assessment is proposed with subjective scores for each sample data. The validity for the designed database has been proved based on the traditional multimedia quality assessment metrics. In addition, an end-to-end 3-D convolutional neural network is introduced to predict the VR video quality without a referenced VR video. This method can extract spatiotemporal features and does not require using hand-crafted features. At the same time, a new score fusion strategy is designed based on the characteristics of the VR video projection process. Taking the pre-processed VR video patches as input, the network captures local spatiotemporal features and gets the score of every patch. Then, the new quality score fusion strategy is applied to get the final score. Such approach shows advanced performance on this database.

INDEX TERMS Virtual reality quality assessment, benchmark database, 3D convolutional neural networks, spatiotemporal features, quality score fusion strategy.

I. INTRODUCTION

With the increase of the variety of multimedia, human beings have more access to receive visual information. As a new simulation and interaction technology, virtual reality (VR) technology is used in many fields [1] such as architecture, military affairs and game. It can create a virtual environment which is consistent with the real world rules, or build a complete hypothetical environment which is contrary to reality. At present, the implementation of VR technology is very challenging. On the one hand, VR requires more complex implementation conditions [2]. People must be equipped with specific devices to feel the immersion [3] of VR. The related equipment and application scenarios restrict the further development of VR. On the other hand, VR requires a variety of perceptual information to match each other to achieve a good quality of experience, and the content of VR is different from the traditional media. Therefore, it is necessary to evaluate the quality of all aspects of VR to promote the more standardized development of the industry. VR video, also known as panoramic stereoscopic video, is a video work played by virtual reality output device. Its purpose is to bring immersive experience with on-the-spot interaction for users to watch videos. Good visual information can bring immersion in the virtual scene, while low quality visual information not only brings bad experience [4], but also can lead to physical disease. As the carrier of VR visual information, VR video requires people to design the appropriate method for virtual reality video quality assessment (VRVQA).

Similar to other multimedia quality assessment (MQA) methods, VRVQA can be devided into two types: subjective assessment and objective assessment [5], [6]. Particularly, subjective evaluation is based on the human observers and objective evaluation provides an index generated by the machine to fit the human observers. That is, an objective assessment based on the algorithm is committed to achieve agreement with subjective results. Similar to video quality assessment (VQA), the objective VRVQA method is divided into three types: full-reference (FR), reduced-reference (RR) and no- reference (NR). FR methods need all the information of the original video. RR methods need some information of the original video. NR methods can obtain the video quality without analyzing any original video information. Therefore, NR methods have more application value and research significance.

In addition, the design of objective evaluation algorithms for multimedia usually requires an authoritative subjective database as a benchmark. In image quality assessment (IQA) and video quality assessment (VQA), there are some famous datasets to measure the validation for the different evaluators: LIVE database [7], [8], IVC database [9], MCL database [10], NBU database [11], NAMA3D database [12] etc. However, there is only one database for video frame rate, bit rate, and resolution in the field of virtual reality assessment [13]. Based on the situation, it is necessary to construct a specially designed dataset for VRVQA, which will be done in this paper.

What needs to be emphasized here is that the VR video acquired by this database belongs to 3D panoramic video, also called 360 degree 3D virtual reality video. In all areas related to quality assessment, the most significant areas are stereoscopic image quality assessment (SIQA) and stereoscopic video quality assessment (SVQA). Similar to stereoscopic video, VR video can be divided into left and right views for stereo perception. Specially, we can not fully utilize the method based on stereoscopic video quality assessment for VRVQA. Because there are many differences between VR video and traditional stereoscopic video [2], [14], [15]. First of all, the filming and production process of VR video is more complicated. In comparison with stereoscopic video, VR video introduces more factors, such as video splicing and synchronization. Secondly, unlike the little angle of stereoscopic video, VR video has a free and all-around perspective which will provide an immersive experience. Therefore, it is necessary to consider extracting more complex features. Finally, VR video's transmission and playback involve the transformation of the plane model and the spherical model. The VR video is in the plane model when it is transmitted, and it is projected into the spherical model when people watch it by the helmet. This process is shown in Fig. 1. So it will bring about more challenges. In order to make readers understand better, We show a VR video frame and compare it with a stereoscopic video frame in Fig. 2. Based on the above discussion, we need to take full account of the complexities of VR video.

In this work, we try to find a VRVQA method that takes full account of VR video Characteristics. In recent years, the method of deep learning has been widely used in the field of multimedia quality assessment. In addition to the most common convolutional neural networks (CNN) models, many other models have implications for their use, such as DNN-based methods [16], [17], methods based on Convolutional Restricted Boltzmann Machines [18], methods based on generating confrontation networks [19]. Considering the factors that affect the quality of VR video are more complex, we decide to use the deep learning model for quality assessment rather than manually extracting features. In fact, VRVQA needs to consider 2D video quality, depth perception, visual comfort, illusion of immersion and other factors. In order to fully consider the VR video's information on the time domain, we decide to design a 3D CNN architecture to capture the spatiotemporal features. We propose a freely available dataset (VRQ-TJU) for VRVQA. A large number of experiments show that our method has achieved good results. In summary, our key contributions are as follows:

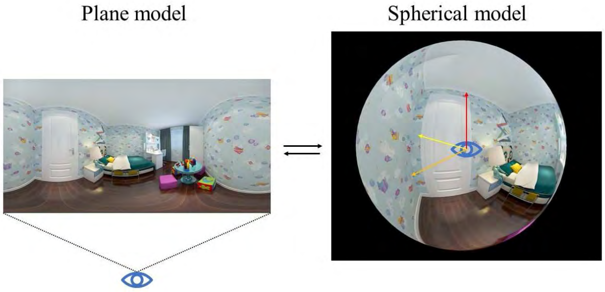


FIGURE 1. **The projection process of VR video. When we watch a VR video of a spherical model, our perspective is at the center of the ball.**

(1) We present an end-to-end 3D CNN based framework for VRVQA, which takes VR difference video patches as input and considers the information among different frames. This is a NR VRVQA method. We can utilize the 3D CNN architecture for quality assessment without the sophisticated preprocessing. To the best of our knowledge, we are the pioneers to exploit the 3D CNN to evaluate the quality of VR video.

(2) We construct a dataset, VRQ-TJU, for the virtual reality quality assessment. If there is no reasonable database as a support, any evaluation algorithm will be meaningless. In essence, the establishment of the database will promote the development of virtual reality evaluation.

(3) We design a quality score fusion strategy for VR plane videos. Different from the spherical model, the spatial distribution of VR videos are uneven in the plane model. This characteristic is due to the specific shooting process and projection process of VR video. Experiments show that the proposed score fusion strategy can effectively improve the performance of quality assessment.

In the following sections, we describe related works (Section II), analyze the data set construction (Section III), explain our proposed method (Section IV), evaluate our method (Section V), draw conclusions and discuss future works (Section VI).

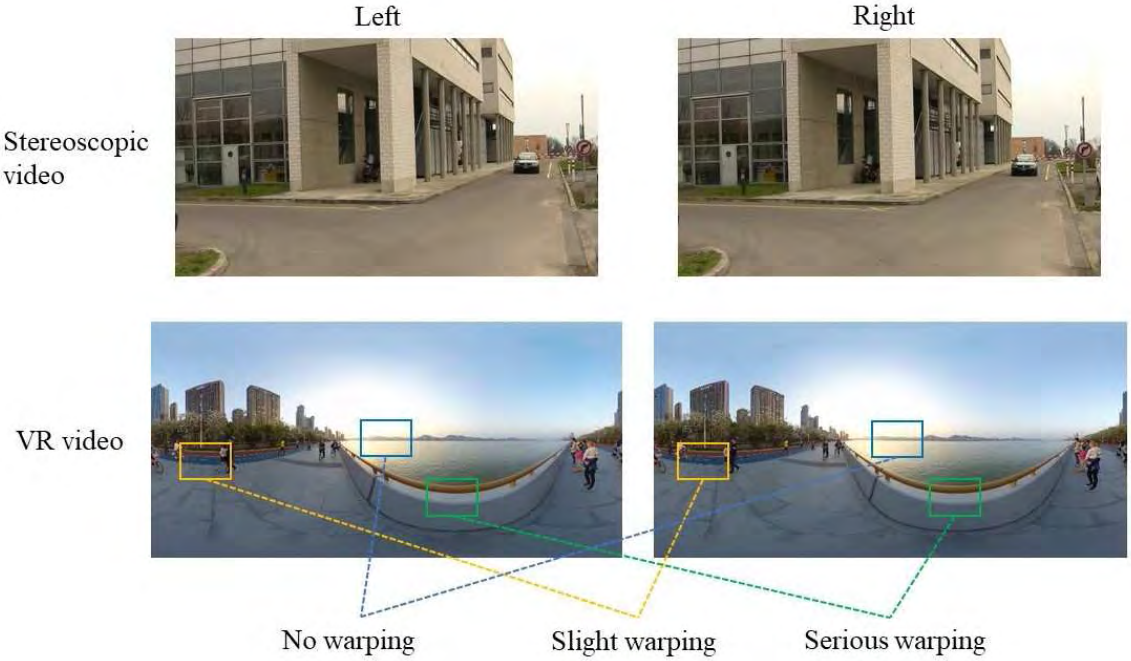


FIGURE 2. One frame in stereoscopic video is compared with one frame in VR video. The same as stereoscopic video, VR video produces stereoscopic perception through the left and right views. However, there are varying degrees of warping in VR video. The greater the vertical distance from the video center, the more warped it is.

II. BACKGROUND ANALYSIS

A. VR VIDEO PRODUCTION AND PROJECTION PROCESS

Unlike ordinary video production processes, VR video production requires a professional camera and a matching post-processing system. First of all, this professional video camera requires multiple directions of video as input to ensure panoramic view and stereoscopic performance. In order to guarantee the complete visual information of all angles, the video in different directions is strictly fixed at the corresponding angle. Then these videos need to be projected the spherical model according to the corresponding angle, such as equi-rectangular projection (ERP) [14]. Finally, multiple videos need to be spliced and blended [20], [21], made distortion correction for each video. In order to ensure that the splicing will not leave any significant traces, we must ensure that when shooting the video in each direction there is redundancy in order to modify [22]. In the splicing and integration process, the synchronization problems between different video has to be considered. Usually it is the technology of genlock from the hardware that is used to solve this problem. We show the process of VR video production in Fig. 3. It should be noted that the VR video we made is in natural scenes.

It was mentioned before that the VR video involved the transformation between a plane model and a spherical model. Such the process is called projection. Most VR videos use the ERP, and the video in our database also adopted this approach. Taking the world map as an example, the central idea of this method is stretching each part of latitude as the length of the equator. Due to the longitude variation of 2pi and latitude variation of pi, such projection video usually presented in a 2:1 ratio of width to height. Polar place is greatly stretched after the process of ERP. Therefore, the space ratio of the poles is pulled in the plane model which produced more redundant pixels. ERP is as shown in Fig. 4. The projection relation from sphere to plane is as follows:

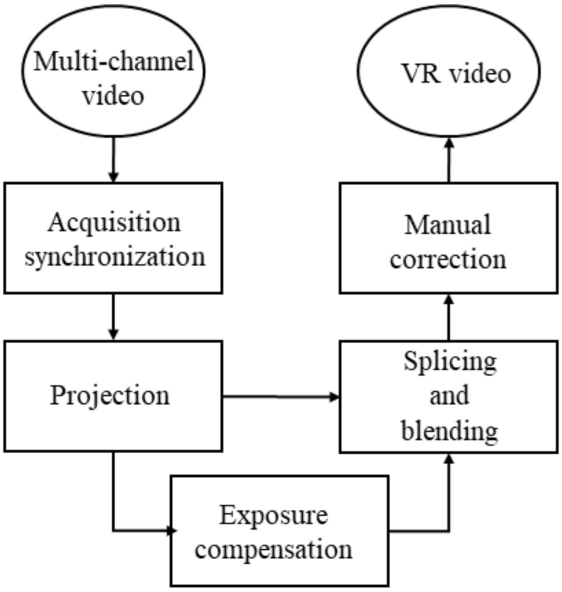


FIGURE 3. The process of VR video production.

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

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

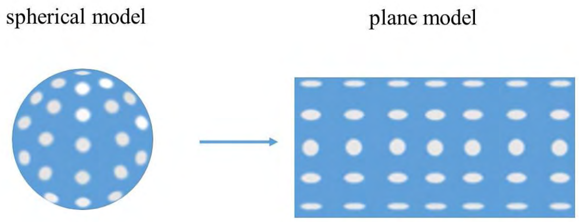


FIGURE 4. The spherical model is projected onto the plane model. The white circle represents varying degrees of warping in the projection process. We notice that the longer the vertical distance from the video center, the more the space is stretched.

In the spherical model, λ denotes longitude, φ denotes latitude. λ_0 denotes central meridian, it can be modified as required. φ_0 denotes standard parallels, which mean the invariant latitude in the projection process. In VR video projection, λ_0 is often equal to 0. In the plane model, x denotes the horizontal coordinate, and y denotes the vertical coordinate.

B. IMAGE AND VIDEO QUALITY ASSESSMENT:DATASETS AND METHODS

1) DATABASES AND METHODS FOR SIQA

Zhou *et al.* [11] established an open stereoscopic image database based on subjective evaluation. In order to reduce the impact of viewing 3D images, Lee *et al.* [23] proposed a subjective experiment of pairing comparisons. On the other hand, the proposed algorithms [24], [25] can more objectively predict the image quality level. Stereoscopic image quality assessment is mainly composed of three categories. Initially, researchers directly used the method of 2D IQA to evaluate the quality of stereoscopic images, such as PSNR, SSIM [26], MS-SSIM [27],GSM [28] and others [29]–[31]. The researchers used the IQA methods for the left and right images of stereo images, and then weighted the scores [32]–[34]. However, these methods do not take into account the depth information in the image. Based on previous work, researchers added depth maps and parallax maps to the assessment criteria. For example, Benoit *et al.* [35] evaluated the distortion of disparity map. Ma *et al.* [36] used Natural Scene Statistics and Structural Degradation to evaluate stereoscopic image quality. Yang *et al.* [37] combined deep learning models and applied SIQA to multimedia analysis towards Internet-of-things. Recently, researchers have combined SIQA with human binocular vision system (HVS) to simulate the attributes of visual perception. Maalouf and Larabi [38] applied a multispectral wavelet decomposition to the two cyclopean color images in order to describe the different channels in the HVS. Ryu and Sohn [39] introduced a model for binocular quality perception.

2) DATABASES AND METHODS FOR SVQA

In [40], authors set up a database based on various packet loss with 2D metrics. Much of SVQA's work draws on the thinking of SIQA. Initially, the researchers also evaluated the left and right views without considering the depth of information. These Methods include PSNR based on method [41], SSIM based on method [40], VQM based on method [42] and so on. On the basis of evaluating color quality, Hewage *et al.* [43] added depth information evaluation. Malekmohamadi *et al.* [44] used the gray level co-occurrence matrix to evaluate the quality of stereoscopic video. In addition to these methods, people have come up with some new metrics for stereoscopic video. Xing *et al.* [45] proposed three perceptual attributes based on the human visual system (HVS), which are shadow degree, separation distance, and spatial position of crosstalk. In [46], a new HVS model with the phenomena of binocular suppression and recurrent excitation was proposed. In general, all of the above approaches rely on manual extraction of features to express part of the characteristics of the assessment object, which is inflexible and time-consuming. Therefore, the depth learning method is expected to evaluate VR video more comprehensively.

C. CONVOLUTIONAL NEURAL NETWORK FOR VISUAL INFORMATION PROCESSING

Nowadays, more and more people choose to use deep networks to handle computer vision-related tasks. As the most commonly used network, CNN is utilized to address detection, classification, tracking and other issues. Recently, CNN also demonstrated its strength in terms of IQA and SIQA. Kim *et al.* [47] made a comprehensive discussion on the field of image quality assessment based on deep learning and focused on the application of CNN network. Kang *et al.* [48] proposed a simple convolutional neural network to predict image quality. This method combines feature learning with regression. Zhang *et al.* [49] proposed two different CNN networks to SIQA. The proposed CNN can learn the local structures which are sensitive to human perception and representative for perceptual quality evaluation. Fan *et al.* [50] performed non-reference IQA based on Multi-expert Convolutional Neural Networks. Compared with the way of extracting features by hand, CNN can learn the mapping relationship between training data and tag more simply and more effectively. Quality assessment scores are also easier to apply by designing an end-to-end CNN.

However, when video is used as training data ([51]–[53]), CNN cannot fully consider the information between adjacent video frames. For this reason, some improvements have to be made to CNN. One solution is to extract the features with temporal information as the input of CNN, but this method is difficult to generalize the quality of video and makes the algorithm more tedious. In order to solve this problem, 3D CNN is used for video field. Ji *et al.* [54] used 3D CNN to capture the motion information of adjacent frames, and then the data with advanced features were regularized and combined with a variety of models to identify human movements. Diba *et al.* [55] also proposed an end to end 3D CNN model to classify videos. The model combined deep motion features into appearance model with optical flow features

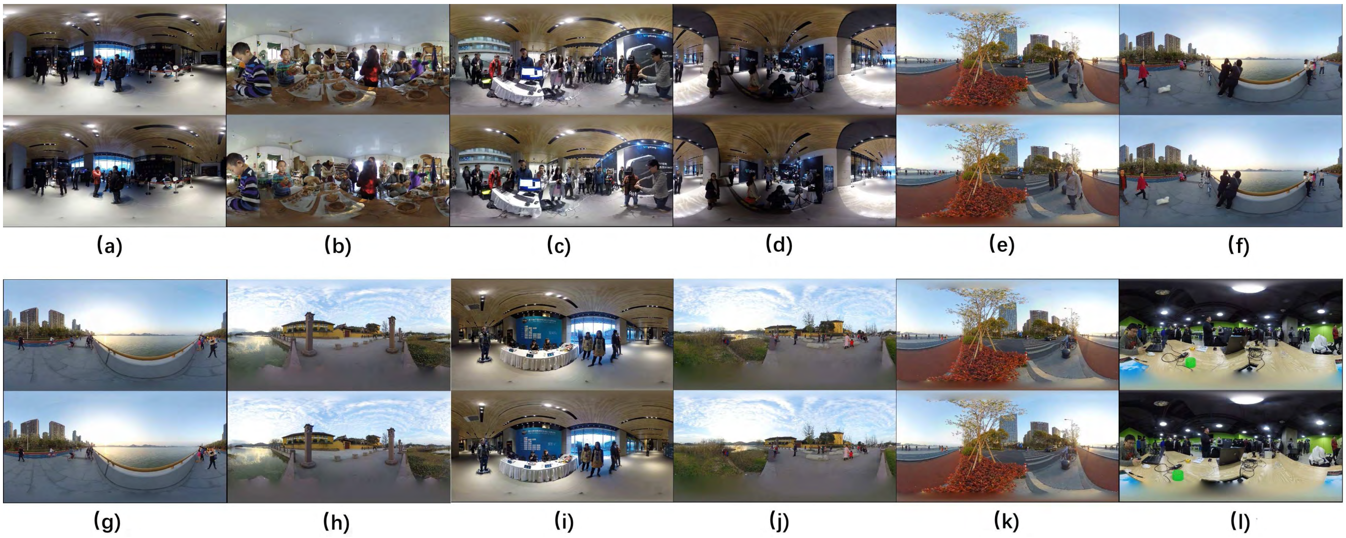


FIGURE 5. 12 original video reference frames in the database(there are altogether 13 original videos, one of which is shown in Figure. 7). (a) Chat. (b) Cook. (c) Demonstartion. (d) Experience. (e) Pedestrien. (f) Photograph. (g) Riverside. (h) Scenic Spot. (i) Sign in. (j) Tourist. (k) Traffic. (i) Wait.

inside the network. In summary, 3D CNN can effectively extract time domain information in VR video. Compared with the traditional method of extracting features manually, 3D CNN can extract higher level and more comprehensive features. Naturally, these features are more suitable for the quality assessment of VR video.

III. VRQ-TJU: DATASET CONSTRUCTION

It is a multi-faceted and complex task to conduct a human study on research of visual multimedia perceptual quality. However, it is necessary and desirable to have a diverse dataset on which the different evaluation performance can be analyzed. In this paper, we build the designed database (VRQ-TJU)[1] towards the VRVQA development. The details on creating process will be given in this section.

A. SEQUENCES ACQUISITION FOR THE SOURCE DATA

As the first step of building a database, how to construct sample structure is the primary problem. Multimedia quality evaluation database building needs to follow some basic principles, and VRVQA is no exception. First of all, the number of lossless benchmark sources should be more than 10. In addition, resolution and frame rate need to be taken into account. Finally, we must give full consideration to the needs of subsequent studies. In addition, we will give a further detailed description on the content, which can also be referred in Table 1.

In Fig. 5, we plot 12 reference VR sources that will be used for the perception. In addition, the sample in Fig. 7 is also included in the VRQ-TJU dataset. So the total referenced VR source number is 13. Here, we have fixed names for each set of benchmark VR: chat, cook, demonstration, experience, field, pedestrian, photograph, riverside, scenic spot, sign in, tourist, traffic and waiting. Resolution is critical to visual perception. In order to provide convenience to the future assessment methods and expand its application value, we keep the sequences resolution in 2560×2560. The particular resolution setting can meet the normal demand. In addition to the spatial domain, changes in the temporal domain also affect the perceived quality of the VR. Just as traditional video display, VR frame rates need to meet certain requirements to match the visual memory of the experiencers. In VQA-TJU, the frame rates are set as 30 fps, and the total frames for each VR source is 507.

TABLE 1. Detail parameters for sequences acquisition of the source data to build VRQ-TJU.

Para	Number
Display Resolution	2560 × 2560
Framerate (f/s)	30
Total frame number for each VR	507
VR source number	13
Symmetric distortion number	104
Asymmetric distortion number	260

B. DISTORTION SIMULATION

The distortions that used in this database are based on the application research in virtual reality. All the degradation in visual quality on each source VR is achieved with a control parameter in a particular range, which is shown in Table 2.

In this paper, H.264 compression and JPEG2000 compression are mainly considered as the distortion factors. Specially, H.264 distortion is simulated using the Vegas media software. JPEG2000 distortion is simulated in Kakadu software.

[1] ftp://eeec.tju.edu.cn/VR (Username:123; Password:123)

TABLE 2. Distortion levels of different video.

Symmetry parameters		Asymmetry parameters							
H.264 (qp)	JPEG2000 (Mb/s)	H.264 (qp)				JPEG2000 (Mb/s)			
Left and Right	Left and Right	Left	Right	Left	Right	Left	Right	Left	Right
30	16	30	0	30	40	16	-	16	64
35	32	35	0	30	45	32	-	16	128
40	64	40	0	35	40	64	-	32	64
45	128	45	0	35	45	128	-	32	128
-	-	30	35	40	45	16	32	64	128

Just as the stereo image quality assessment, VR should be considered in symmetrical design and asymmetrical design. For stereoscopic signals, there are some research activities in asymmetric compression [32], [56]. However, the visual quality based on asymmetric distortions should be pushed as an interesting avenue. In the designed database, we applied the different distortion degree in left image and right image.

In total, 104 pieces of virtual reality in symmetric distortion and 260 pieces of virtual reality in asymmetric distortion are obtained. Based on the acquired data, the subjective study can be made on this project.

C. SUBJECTIVE EVALUATION

In order to make a complete quality evaluation in single-stimulus with hidden reference study [6], the study was conducted at Tianjin University, which took about 20 weekly days. First of all, the subjective evaluation team with 30 people is set up, which takes into account many subjective evaluation principles. Specifically, the number of male and female in the evaluation team is the same. And it also includes a number of non professionals. Before the subjective evaluation, we conducted verbal instructions, to make participants more clear in visual content evaluation [57]. In order to prevent subject fatigue, the evaluation time for each participate should be less than 30 minutes. Of course, the subjective assessment in our work is consistent with the principles.

In the process, two major issues need to be specifically addressed: playing equipment and number of participants. Based on the exist recommendations, some researchers in this field have published some study results on the maximum number for subjects. It is believed that enough subjects can be a more accurate metric. This paper designs that each VR should be rated by 30 participates. Based on [57], we can get the averaged across subjects into mean opinion scores (MOS). Particularly, the subject rejection was used for recommendations, in which two subjects scores were rejected [57].

The designed database consists of 13 referenced VR, 104 pieces of virtual reality in symmetric distortion and 260 pieces of virtual reality in asymmetric distortion are with associated MOS. A histogram of MOS is shown in Fig. 6a. The standard deviation of subjective ratings is an important measure of the degree of dispersion of subjective ratings. For different types of samples, we calculate the standard deviation of the subjective ratings of each type of VR video as $\sigma\{\mathrm{Sym}, \mathrm{Asym}, \mathrm{H.264}, \mathrm{JPEG2000}\} = \{0.39, 0.47, 0.45, 0.44\}$. We note that these standard deviations are in line with previous studies of this nature for images and videos [6], [7], [58]. Further, the MOS distribution is uniform through a large portion of the scale indicating that the distortions in the VRQ-TJU database span a wide range of visual quality.

D. PRIMARY EVALUATION BASED ON TRADITIONAL MULTIMEDIA QUALITY ASSESSMENT METRICS

As the two most basic evaluation indicators of multimedia data, structural similarity (SSIM) and peak signal to noise ratio (PSNR) are widely used in primary evaluation for the subjective database building.

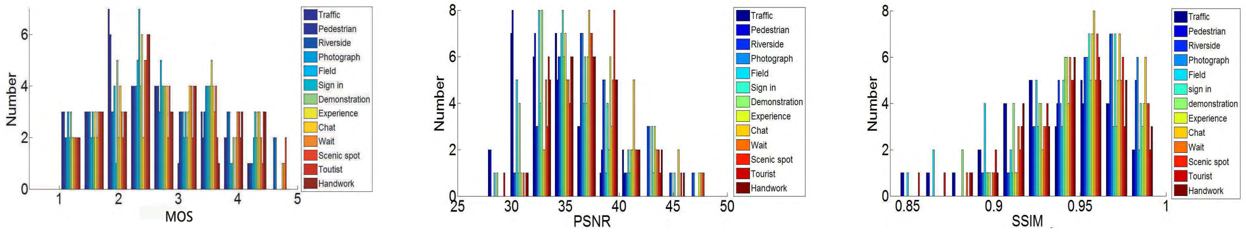
For the newly created VRQ-TJU database, the SSIM values and the PSNR values of all samples are computed. And its histogram distribution are shown in Fig. 6b. and Fig. 6c. According to this situation, the distributional rationality is judgable.

IV. PROPOSED METHOD

We first perform a simple preprocessing of the video and then utilize an end-to-end 3D CNN to get the local quality score containing the space-time information. Finally, we use a quality score fusion strategy to get the overall video score. This section is divided into three parts to introduce our method in detail.

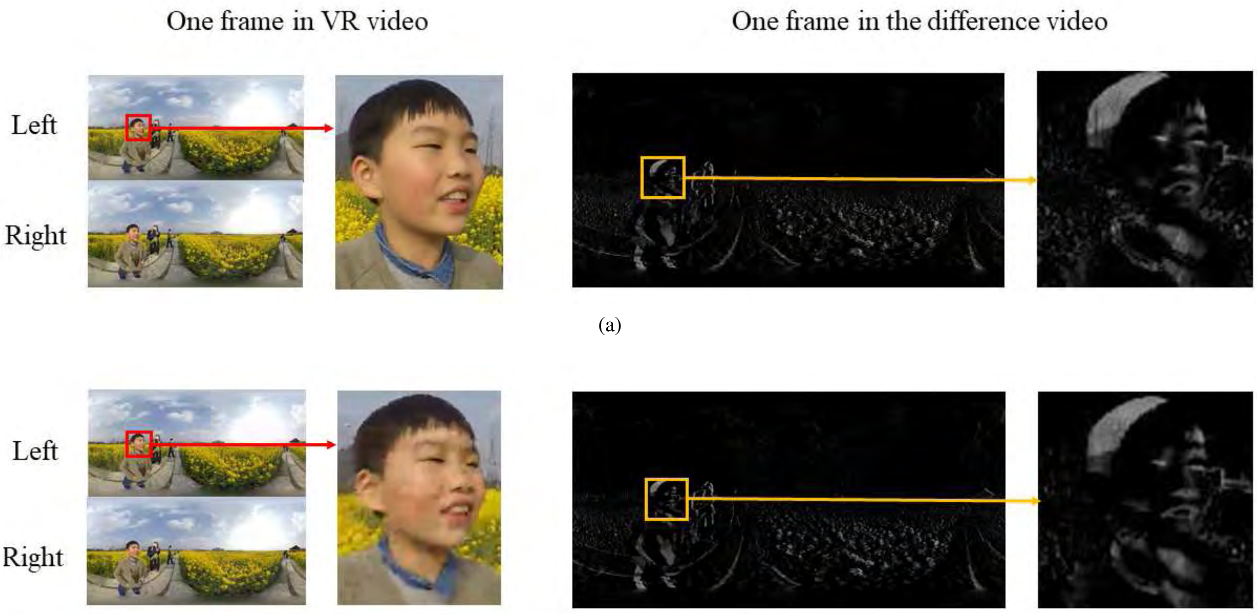
A. DATA PROCESSING

In the plane model, the VR video consists of two 2D panoramic videos with parallax. Two 2D panoramic videos respectively correspond to the left view and the right view. In VRVQA, the depth perception information and stereo perception information must be considered between the two videos. The two types of information are essential to improve the immersion of VR video. Based on the following two reasons, we decide to extract data from the difference video and then put it into 3D CNN. On the one hand, difference video



(a) (b) (c)

FIGURE 6. The histogram statistical results of subjective MOS, PSNR and SSIM.



(a)

(b)

FIGURE 7. Difference frames from VR video frames. (a) 50th frame of a reference VR video and the corresponding difference frame, MOS=4.90. (b) 50th frame of a distored VR video and the corresponding difference frame, MOS = 1.20.

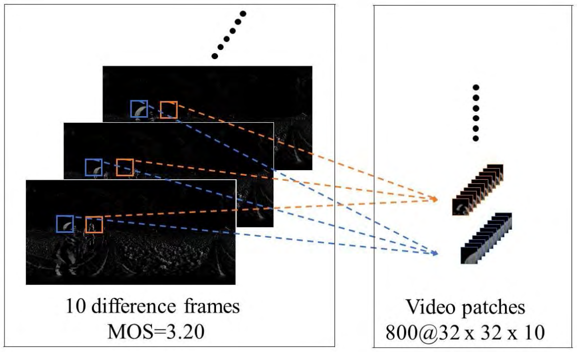


FIGURE 8. The process of splitting VR video.

is mainly for the depth perception information and stereo perception information. This type is exactly what we need. Zhang *et al.* [49] and Ma *et al.* [59] concluded that the difference image is more valuable than the left and right views in SIQA. In our previous work [60]–[62], the difference image calculated from left and right views has been demonstrated to retain stereoscopic perception information, which can be used to represent the quality of stereoscopic image. There is no difference between VR video and stereoscopic video in this respect. On the other hand, because VR videos contain panoramic view, complex preprocessing of VR video will bring more challenges. To make our approach more practical and efficient, we decide to get the data from the difference video. Firstly, extracted video frames are transformed into gray image. We calculate the value of the difference video V_d at position (x, y, z) as

$$V_d(x, y, z) = |V_l(x, y, z) - V_r(x, y, z)| \tag{3}$$

where V_l and V_r are the left and right view frames. In order to demonstrate this concept more intuitively, several samples are enumerated in Fig. 7. The VR video is projected as a spherical model when viewed, with the view of the viewer focused only on one part of it. This is equivalent to VR video being magnified when viewed, so small changes in the plane model can greatly affect the normal viewing rating.

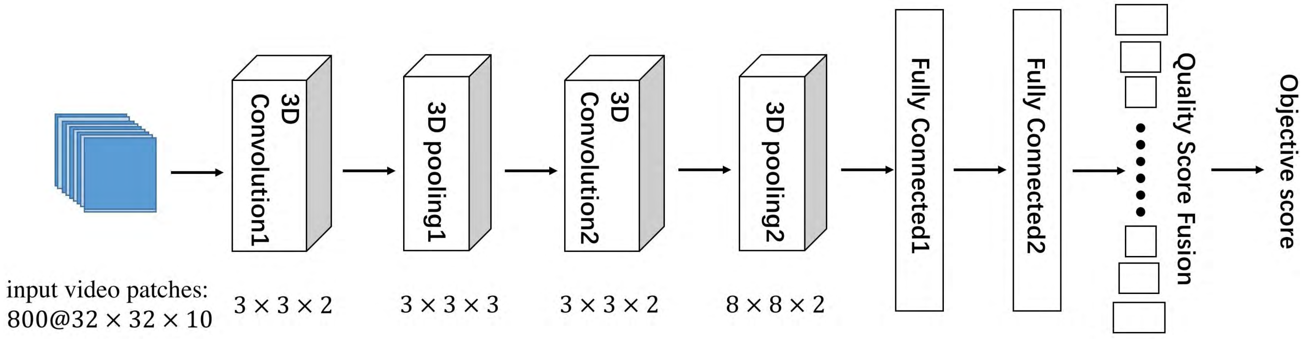


FIGURE 9. **The end to end 3D CNN architecture. Each video has 800 video patches as input.**

Generally, the use of deep network to obtain satisfactory indicators requires a large amount of labeled data. However, due to the scale of VR video data set, if the data is not enhanced, the network model will not get a good generalization performance. At the same time, compared with the spherical model, the spatial distribution of VR video is uneven in the plane model. Therefore, it is necessary to evaluate the local video first and then to evaluate the quality of the video as a whole. We will elaborate the reason later. Taking into account these two points, we decide to split VR video to increase the size of the sample data. A lot of small VR video patches are obtained by splitting video in time and space dimensions. Specifically, ten frames are taken evenly from each video, and are divided into the image patches with the resolution of 32 × 32 at the same position in each frame. Ten image patches constitute a 32 × 32 × 10 video patch as input. Then the patch is marked with the subjective score of the video that contains the patch. By splitting the VR video without overlapping, each of the database's VR videos can get 800 video patches. The specific process is shown in Fig. 8. So this way effectively improves the sample size, satisfies the needs of 3D CNN, and provides a new thought for the quality score fusion strategy. Thus better adapt to the characteristics of VR video.

B. 3D CNN

The end to end 3D CNN architecture is shown in Fig. 9. Our 3D CNN architecture consists of two 3D convolution layers, C1, C2, two 3D pooling layers, S1, S2, and two fully-connected layers, FC1, FC2. In this section, the structure of 3D CNN is described in order.

1) 3D CONVOLUTION

In 2D CNN, convolution can only express two-dimensional feature maps. Therefore, The information on the time domain will be lost after every 2D conolution operation. Compared with 2D convolution, 3D convolution preserves the input time information, and is more suitable for video analysis. 3D convolution can be understood as convolution of adjacent multiple frame images with multiple different convolution kernels. All convolution results are summed to obtain the feature mapping. The convolution process is actually a linear operation between the input data and the kernel function. The calculation of 3D convolution can be expressed as

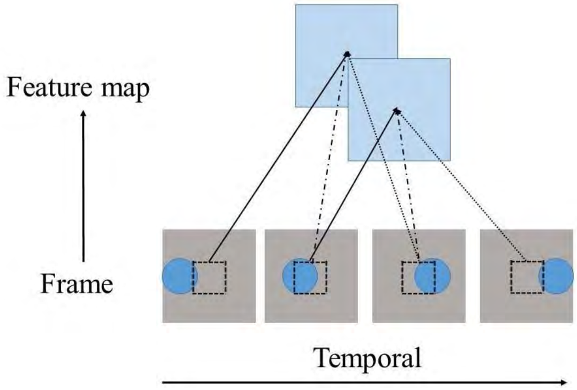


FIGURE 10. **In 3D convolution, each location of the 3D feature maps obtained by the convolution of the same location in several adjacent input frames and 3D kernels. The shared weights are in the same kind of straight line. So the time domain feature of the video can be extracted.**

$$Conv(x, y, z) = AF(b + \sum_p \sum_q \sum_r h(x+p, y+q, z+r) \cdot W_{pqr}) \quad (4)$$

AF is on behalf of the activation function, such as Sigmoid, Tanh and ReLU. The ReLU is used for the activation function to increase the nonlinearity after each convolution layer. The b is the bias for the feature map. W_{pqr} represents the parameter of the convolutional kernel. Where $h(x+p, y+q, z+r)$ stands for the input pixel value at position $(x+p, y+q)$ in the $(z+r)$th frame. This paper uses a 3 × 3 × 2 convolution kernel. The process of 3D convolution is shown in Fig. 10. The ReLU expression is as follow:

$$ReLU(x, y, z) = max(0, Conv(x, y, z)) \quad (5)$$

2) 3D POOLING

3D pooling is the process of aggregating the statistics. These aggregated statistics not only have much lower dimensions, but also effectively prevent over-fitting. Take the 3D maximum pooling used in this paper as an example. If a $3 \times 3 \times 3$ max-pooling operator is performed on a $W \times H \times T$ feature map, we collect the max value in each $3 \times 3 \times 3$ non-overlap regions which form a new feature map with size of $\frac{W}{3} \times \frac{H}{3} \times \frac{T}{3}$. The formula for 3D maximum pooling is as follows:

$$Pool(x, y, z) = max(ReLU(x, y, z)) \tag{6}$$

where $Pool(x, y, z)$ represents the feature value obtained after pooling.

3) ARCHITECTURE ANALYSIS

Take a video input patch for example. Video patch size is $32 \times 32 \times 10$. In the C1 with a kernel size of $3 \times 3 \times 2$, the output is 3D feature map whose size is $30 \times 30 \times 9$. Then the 3D feature map through the S1 with a kernel size of $3 \times 3 \times 2$ and get the 3D feature map which size is $10 \times 10 \times 3$. Next into the C2 whose is the same as the C1. So the output is $8 \times 8 \times 2$ 3D feature map. Then the 3D feature map through the S2 with a kernel size of $8 \times 8 \times 2$. Finally, 3D feature map passes through FC1 and FC2. The output is a 512-D feature vector and the final score in turn. The specific parameters are shown in Table 3.

TABLE 3. The detailed parameters of our proposed method .

Layer	Kernel Size	Output Size	Feature maps
Input	-	$32 \times 32 \times 10$	1
C1	$3 \times 3 \times 2$	$30 \times 30 \times 9$	50
S1	$3 \times 3 \times 3$	$10 \times 10 \times 3$	50
C2	$3 \times 3 \times 2$	$8 \times 8 \times 2$	50
S2	$8 \times 8 \times 2$	$1 \times 1 \times 1$	50

We use the Stochastic Gradient Descent (SGD) algorithm in 3D CNN, which can optimize parameters. The learning rate is initialized to 0.001. An objective function is adopted as follows:

$$min_\theta \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - y_i)^2 + \lambda ||\theta||_F^2 \tag{7}$$

where $f(x_i)$ and y_i denote predicted score and ground-truth score. λ is the regularization parameter. The linear activation function is used at the output. In order to avoid over-fitting, the dropout strategy is used, with parameter of 0.5 after each pooling layer. In the first full connection layer, we use the dropout strategy with parameter of 0.25. The principles of dropout are as follows:

$$\widetilde{y} = P(0, 1) \cdot y \tag{8}$$

$P(0, 1)$ represents a random generation of 0 or 1. y is the input of the dropout layer, $\widetilde{y}$ is the output of the dropout layer. For the sample data, 60% were used for training, 20% were used for validating and 20% were used for testing.

C. QUALITY SCORE FUSION STRATEGY

Because the spatial distribution of VR video becomes uneven during projection, a new scoring strategy needs to be developed to match this characteristic. By 3D CNN, evenly distributed and non-overlapping video patches are numerously acquired. Each video patch has an objective score. It is important to emphasize that when the network is trained, the fraction of the patch is based on the subjective perception of the whole video in the spherical model. Because these video patches are based on the plane model, the closer to the poles, the smaller proportion of the patch in the actual viewing space. Therefore, a score fusion strategy is designed as follows:

$$S_f = (\sum_x \sum_y S_{xy} W_{xy}) / \sum_x \sum_y W_{xy} \tag{9}$$

$$W = cos(\pi \frac{h'}{h}) \tag{10}$$

S_f denotes the final score, S_{xy} denotes the objective fraction of all video patches in the VR video. The objective fraction is multiplied, of each video patch by the weighted W_{xy}. h' denotes the vertical distance between the center point of the video patch and the whole video center. h denotes the vertical height of VR video. It is worth noting that the value of the weight W is only related to the vertical distance between center position of video patch and the central position of VR video. This feature is determined by the VR video projection process.

V. EXPERIMENT

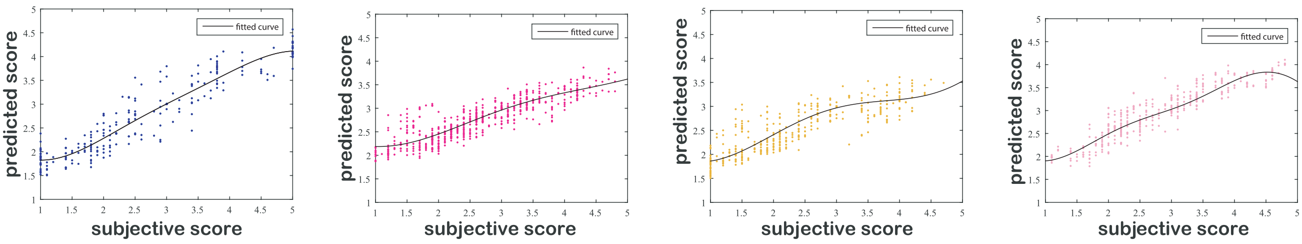
In this section, the efficiency of our algorithm will be evaluated and our approach will be compared with some traditional MQA methods in VRQ-TJU.

A. COMPARISON OF 3D CNN PARAMETERS

We select some parameters and design a contrast experiment to choose the best network. In addition to the parameters used in the article, we also use two comparison values. These parameters include learning rate (lr), epoch (ep), batch size (bs) and number of convolution cores (nocc). It needs to be explained that the parameters of the final model are as follows. The learning rate is 0.001, epoch is 200, batch size is 128, number of convolution cores is 50. The specific parameters are shown in Table 4.

B. COMPUTATIONAL ENVIRONMENT AND COST

The proposed 3D CNN model are developed by the Python deep learning library Keras which uses the Theano backend. Hardware configuration has a single 3.2GHz CPU and a single GTX1080 GPU. The sample size has a total of 301,600 video patches with a size of $32 \times 32 \times 10$. Our 3D CNN model only consists of six layers. For the entire database, our model's training time is 3.61 hours. In other words, the running time of a video is 43.03 seconds, and a video patch has a runtime of 0.054 seconds. This index shows that our method is efficient. The training obtained model can



(a) (b) (c) (d)

FIGURE 11. Predicted MOS vesus subjective MOS. (a) the sample based on symmetric distortion. (b) the sample based on the asymmetric distortion. (c) the sample based on the H.264 distortion type. (d) the sample based on the JPEG2000 distortion.

TABLE 4. Comparison of 3D CNN parameters.

	PLCC	SRCC	KROCC	RSME
lr = 0.01	0.8064	0.8006	0.6681	11.7936
lr = 0.005	0.8882	0.8816	0.7926	8.5463
ep = 120	0.8998	0.8929	0.7991	8.2225
ep = 160	0.9001	0.8933	0.8031	8.2219
bs = 64	0.9006	0.8937	0.8098	8.2176
bs = 256	0.9007	0.8939	0.8105	8.2214
nocc = 32	0.8943	0.8872	0.7936	8.1978
nocc = 64	0.9007	0.8939	0.8117	8.2154
Final model	0.9008	0.8940	0.8118	8.1985

directly obtain the objective score of the corresponding video patch, so there is almost no time cost. At the same time, our method does not need to refer to the original VR video, The above two points show that our algorithm has very good practical value.

C. OVERALL PERFORMANCE

This evaluation uses four indicators, which are Pearson linear correlation coefficient (PLCC), Spearman rank order correlation coefficient (SROCC), Kendall rank-order correlation coefficient (KROCC) and Root mean squared error (RMSE).

In order to be more comprehensive in response to the experimental results, five different experiments will be performed: experiment based on all samples, experiment based on symmetrical distortion samples, experiment based on asymmetrical distortion samples, experiment based on JPEG2000 distortion samples, experiment based on H.264 distortion samples. And the PLCC, SRCC KRCC and RMSE are shown in Table 5. As can be seen from the data, the four indexes are well represented in the whole experiment.

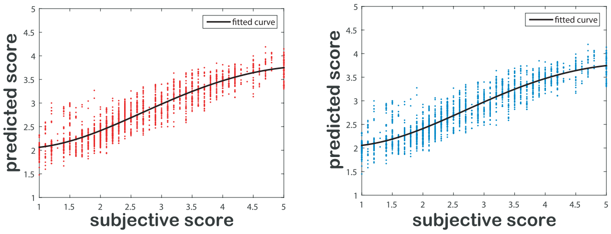
In order to observe the methods consistency with the subjective feeling more intuitively, the scatter diagram is chosen to do it. In Fig. 11, it shows the four different pairs of data based on the subjective perception and objective scores.

The method of averaging the objective score of all VR video patches, is named the "3D CNN+AVG". The 3D CNN method which uses the quality score fusion strategy is named "3D CNN+QSFS". In Fig. 12, the scatter plot is shown between the predicted score and the objective score of 3D CNN+AVG and 3D CNN+QSFS. These scatter plots illustrate that the subjective score has a good linear correlation with the objective score.

In order to ensure the reliability of the experimental results, we conduct the train-test process 30 times and calculate the average value of all the results as the final indicator. In Fig. 13,

TABLE 5. Performance on the whole dataset.

Para	PLCC	SRCC	KROCC	RSME
Whole database	0.9008	0.8940	0.8118	8.1985
Symmetrical distortion	0.9264	0.9033	0.8309	7.8136
Asymmetrical distortion	0.8688	0.8438	0.7726	7.7733
H.264 distortion	0.8685	0.8642	0.7806	8.8136
JPEG2000 distortion	0.9183	0.9102	0.8425	7.5733



(a) (b)

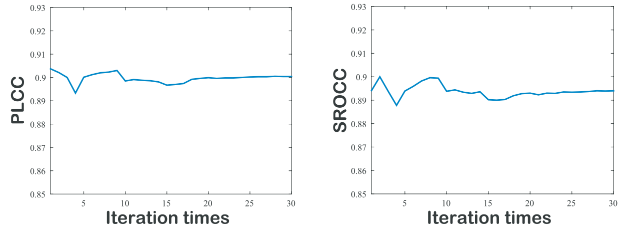
FIGURE 12. Predicted MOS vesus subjective MOS. (a) the scatter plot of 3D CNN+AVG. (b) the scatter plot of 3D CNN+QSFS.

TABLE 6. Performance comparison with traditional MQA methods.

Metrics	PLCC	SROCC	KROCC	RMSE
PSNR	0.7949	0.7974	0.6135	11.4875
SSIM [26]	0.8057	0.8278	0.6483	12.3420
the method of IQA [28]	0.7586	0.7236	0.6134	14.2434
the method of SIQA [35]	0.8134	0.8048	0.7986	10.4853
the method of VQA [6]	0.8246	0.8085	0.7864	9.3975
the method of SVQA [63]	0.8317	0.8196	0.8034	9.3552
3D CNN+AVG	0.8893	0.8823	0.8054	8.2138
3D CNN+QSFS	**0.9008**	**0.8940**	**0.8118**	**8.1985**

TABLE 7. Performance comparison with traditional MQA methods on each sample.

	Symmetry parameters				Asymmetry parameters			
Metrics	PLCC	SROCC	KROCC	RMSE	PLCC	SROCC	KROCC	RMSE
PSNR	0.8450	0.8572	0.6682	14.2443	0.7721	0.7672	0.5862	11.3432
SSIM [26]	0.8660	0.8835	0.7055	12.3432	0.7640	0.8054	0.6287	13.2332
the method of IQA [28]	0.7541	0.7243	0.7029	13.5644	0.7573	0.7261	0.7184	13.5432
the method of SIQA [35]	0.8523	0.8423	0.8044	13.5583	0.7843	0.7638	0.7467	14.3459
the method of VQA [6]	0.8583	0.8351	0.8245	12.4583	0.8023	0.8124	0.7547	11.4324
the method of SVQA [63]	0.8553	0.8486	0.8344	11.3334	0.7943	0.7861	0.7439	10.4334
3D CNN+AVG	0.9159	0.0.8944	0.8217	7.8229	0.8574	0.8321	0.7637	7.7842
3D CNN+QSFS	**0.9264**	**0.9033**	**0.8309**	**7.8136**	**0.8688**	**0.8438**	**0.7726**	**7.7733**
	H.264 distortion				JPEG2000 distortion			
Metrics	PLCC	SROCC	KROCC	RMSE	PLCC	SROCC	KROCC	RMSE
PSNR	0.7814	0.7750	0.6013	11.3443	0.7236	0.7153	0.5489	10.7655
SSIM [26]	0.7464	0.7626	0.6938	10.3843	0.7256	0.8025	0.6282	9.3865
the method of IQA [28]	0.7524	0.7339	0.6937	10.3836	0.7032	0.7244	0.5763	8.9752
the method of SIQA [35]	0.7933	0.8021	0.8194	9.3975	0.7245	0.7043	0.6723	9.3864
the method of VQA [6]	0.7824	0.7242	0.6944	8.8842	0.7284	0.6986	0.7033	9.3862
the method of SVQA [63]	0.8134	0.8061	0.7974	9.2865	0.7344	0.7144	0.6973	8.2946
3D CNN+AVG	0.8591	0.8586	0.7719	8.8035	0.9096	0.9061	0.8338	7.4675
3D CNN+QSFS	**0.8685**	**0.8642**	**0.7806**	**8.8136**	**0.9183**	**0.9102**	**0.8425**	**7.5733**



(a) (b)

FIGURE 13. (a) the mean value of PLCC after different times of iterative. (b) the mean value of SROCC different times of iterative.

taking PLCC and SROCC as examples, the index is plotted, of the proposed method after a number of iterations. At the same time, each retraining will reset the weight of the model, so there is a slight difference in the indicators obtained from each iteration. Here, we need to explain the degree of dispersion of different indicators in all iterations, so as to prove the stability and scientificity of the model. The range of the four indicators evaluated by this method is {PLCC, SROCC, KROCC, RMSE} = {0.0108, 0.0128, 0.0109, 0.1781}

D. DATA COMPARISON WITH THE TRADITIONAL METHOD

In order to make the data more convincing, this paper chooses 6 typical methods of MQA as comparison. PSNR and SSIM are considered first as a classic method for comparison. The method proposed in the literature [28] represents the method of IQA. The method proposed in the literature [35] was chosen as the representative method of SIQA methods. The method proposed in the literature [6] represents the method of VQA. The method proposed in the literature [63] represents the method of SVQA. The reason for choosing these methods is that by contrast, the pertinence of our method can be better highlighted. In order to prove the effectiveness of our proposed fractional fusion strategy, a contrast experiment is added. Detailed results are shown in Table 6. Black numbers represent the best indicators. Compared with the method of SVQA, the index of 3D CNN+QSFS increased by 0.0691, 0.0744, 0.0084, 1.1567 on PLCC, SROCC, KROCC and RMSE. Compared with 3D CNN+AVG, the index of 3D CNN+QSFS increased by 0.0115, 0.0117, 0.0064, 0.0153 on PLCC, SROCC, KROCC and RMSE. These experimental results show that our proposed method of 3D CNN can effectively extract the local spatiotemporal information in VR video. Our quality score fusion strategy can also further improve the performance of the algorithm.

The data demonstrate that our method is more suitable for VR video than other methods. The best of these six contrasting methods is the method of SVQA. This result is in line with our guess. VR video has the same left and right views as stereoscopic video. Therefore, the evaluation of stereoscopic perception is also applicable to VRVQA. However, these methods do not take into account the unique features of VR video, such as immersion, production processes and playback methods. So these contrasting methods cannot be fully applied to VR video.

We test the four types of samples with 3D CNN and other traditional MQA methods, and then list their indicators in table 7. The experimental results also show that the method proposed by us has a good correlation for different kinds of samples.

VI. CONCLUSION AND FUTURE WORK

In this paper, a NR quality assessment method for VR video was proposed. This work will help VR technology develop more mature. Before this, no one has assessed the quality of VR video and used 3D CNN to the field of quality assessment. We propose a method of using 3D CNN to assess the quality of VR video. Unlike traditional NR-VQA methods, this method does not require complex preprocessing or hand-crafted features. The objective prediction score of the VR video is obtained by the combination of the local spatiotemporal features and the quality score fusion strategy. Experiments show that the results of the algorithm are consistent with the subjective quality assessment.

In the future work, it is expected to use more advanced networks to extract more local spatio-temporal features. At the same time, we hope to combine saliency and projection distortion analysis in future work, not just spatial distribution changes. In addition, it will be continuous to explore the characteristics of VR video to design more targeted algorithms. Finally, we will consider improving the richness of the experiment.

REFERENCES

[1] C. J. Turner, W. Hutabarat, J. Oyekan, and A. Tiwari, "Discrete event simulation and virtual reality use in industry: New opportunities and future trends," *IEEE Trans. Human–Mach. Syst.*, vol. 46, no. 6, pp. 882–894, Dec. 2016.

[2] S. E. Chen, "QuickTime VR: An image-based approach to virtual environment navigation," in *Proc. Conf. Comput. Graph. Interact. Techn.*, 1995, pp. 29–38.

[3] P. Howard-Jones, M. Ott, T. van Leeuwen, and B. De Smedt, "The potential relevance of cognitive neuroscience for the development and use of technology-enhanced learning," *Learn. Media Technol.*, vol. 40, no. 2, pp. 131–151, 2015.

[4] L. Freina and M. Ott, "A literature review on immersive virtual reality in education: State of the art and perspectives," in *Proc. Elearn. Softw. Educ.*, 2015, pp. 133–141.

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

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

[7] A. K. Moorthy, C.-C. Su, A. Mittal, and A. C. Bovik, "Subjective evaluation of stereoscopic image quality," *Signal Process., Image Commun.*, vol. 28, no. 8, pp. 870–883, Dec. 2013.

[8] M.-J. Chen, C.-C. Su, D.-K. Kwon, L. K. Cormack, and A. C. Bovik, "Full-reference quality assessment of stereopairs accounting for rivalry," *Signal Process., Image Commun.*, vol. 28, no. 9, pp. 1143–1155, 2013.

[9] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Quality assessment of stereoscopic images," *Eurasip J. Image Video Process.*, vol. 2008, no. 1, pp. 1–13, 2009.

[10] R. Song, H. Ko, and C.-C. J. Kuo, "MCL-3D: A database for stereoscopic image quality assessment using 2D-image-plus-depth source," *J. Inf. Sci. Eng.*, vol. 31, no. 5, pp. 1593–1611, 2014.

[11] J. Zhou *et al.*, "Subjective quality analyses of stereoscopic images in 3DTV system," in *Proc. Vis. Commun. Image Process.*, Nov. 2011, pp. 1–4.

[12] M. Urvoy *et al.*, "NAMA3DS1-COSPAD1: Subjective video quality assessment database on coding conditions introducing freely available high quality 3D stereoscopic sequences," in *Proc. 4th Int. Workshop Qual. Multimedia Exp.*, Jul. 2012, pp. 109–114.

[13] H. Duan, G. Zhai, X. Yang, D. Li, and W. Zhu, "IVQAD 2017: An immersive video quality assessment database," in *Proc. Int. Conf. Syst., Signals Image Process.*, May 2017, pp. 1–5.

[14] M. Hosseini and V. Swaminathan. (2016). "Adaptive 360 VR video streaming: Divide and conquer!" [Online]. Available: https://arxiv.org/abs/1609.08729

[15] C. Ozcinar, A. De Abreu, S. Knorr, and A. Smolic, "Estimation of optimal encoding ladders for tiled 360° VR video in adaptive streaming systems," in *Proc. IEEE Int. Symp. Multimedia*, Dec. 2017, pp. 45–52.

[16] K. Gu, G. Zhai, X. Yang, and W. Zhang, "Deep learning network for blind image quality assessment," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2014, pp. 511–515.

[17] F. Gao, Y. Wang, P. Li, M. Tan, J. Yu, and Y. Zhu, "DeepSim: Deep similarity for image quality assessment," *Neurocomputing*, vol. 257, pp. 104–114, Sep. 2017.

[18] K. Wang, J. Zhou, N. Liu, and X. Gu, "Stereoscopic images quality assessment based on deep learning," in *Proc. Vis. Commun. Image Process.*, Nov. 2016, pp. 1–4.

[19] F. Gao, J. Yu, S. Zhu, Q. Huang, and Q. Tian, "Blind image quality prediction by exploiting multi-level deep representations," *Pattern Recognit.*, vol. 81, pp. 432–442, Sep. 2018.

[20] M. Brown and D. G. Lowe, "Automatic panoramic image stitching using invariant features," *Int. J. Comput. Vis.*, vol. 74, no. 1, pp. 59–73, Aug. 2007.

[21] R. Szeliski, *Computer Vision: Algorithms and Applications*. New York, NY, USA: Springer-Verlag, 2010.

[22] W. Xu, "Panoramic video stitching," Ph.D. dissertation, Dept. Comput. Sci., Univ. Colorado Boulder, Boulder, CO, USA, 2012.

[23] J.-S. Lee, L. Goldmann, and T. Ebrahimi, "Paired comparison-based subjective quality assessment of stereoscopic images," *Multimedia Tools Appl.*, vol. 67, no. 1, pp. 31–48, Nov. 2013.

[24] F. Battisti, E. Bosc, M. Carli, P. Le Callet, and S. Perugia, "Objective image quality assessment of 3D synthesized views," *Signal Process., Image Commun.*, vol. 30, pp. 78–88, Jan. 2015.

[25] X. Wang, Q. Liu, R. Wang, and Z. Chen, "Natural image statistics based 3D reduced reference image quality assessment in contourlet domain," *Neurocomputing*, vol. 151, pp. 683–691, Mar. 2015.

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

[27] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multi-scale structural similarity for image quality assessment," in *Proc. 37th Asilomar Conf. Signals, Syst. Comput.*, Nov. 2003, pp. 1398–1402.

[28] A. Liu, W. Lin, and M. Narwaria, "Image quality assessment based on gradient similarity," *IEEE Trans. Image Process.*, vol. 21, no. 4, pp. 1500–1512, Apr. 2012.

[29] X. Gao, F. Gao, D. Tao, and X. Li, "Universal blind image quality assessment metrics via natural scene statistics and multiple kernel learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 24, no. 12, pp. 2013–2026, Dec. 2013.

[30] X. Li, D. Tao, X. Gao, and W. Lu, "A natural image quality evaluation metric," *Signal Process.*, vol. 89, no. 4, pp. 548–555, Apr. 2009.

[31] F. Gao and J. Yu, *Biologically Inspired Image Quality Assessment*. Amsterdam, The Netherlands: North Holland, 2016.

[32] P. Gorley and N. Holliman, "Stereoscopic image quality metrics and compression," *Proc. SPIE*, vol. 6803, pp. 680305-1–680305-12, 2008.

[33] P. Campisi, P. Le Callet, and E. Marini, "Stereoscopic images quality assessment," in *Proc. Eur. Signal Process. Conf.*, Sep. 2015, pp. 2110–2114.

[34] J. Yang, C. Hou, Y. Zhou, Z. Zhang, and J. Guo, "Objective quality assessment method of stereo images," in *Proc. 3DTV Conf., True Vis.-Capture, Transmiss. Display 3D Video*, May 2009, pp. 1–4.

[35] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Using disparity for quality assessment of stereoscopic images," in *Proc. IEEE Int. Conf. Image Process.*, Oct. 2008, pp. 389–392.

[36] J. Ma, P. An, L. Shen, and K. Li, "Reduced-reference stereoscopic image quality assessment using natural scene statistics and structural degradation," *IEEE Access*, vol. 6, pp. 2768–2780, 2017.

[37] J. Yang, B. Jiang, H. Song, X. Yang, W. Lu, and H. Liu, "No-reference stereoimage quality assessment for multimedia analysis towards Internet-of-Things," *IEEE Access*, vol. 6, pp. 7631–7640, 2018.

[38] A. Maalouf and M.-C. Larabi, "CYCLOP: A stereo color image quality assessment metric," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process.*, May 2011, pp. 1161–1164.

[39] S. Ryu and K. Sohn, "No-reference quality assessment for stereoscopic images based on binocular quality perception," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 4, pp. 591–602, Apr. 2014.

[40] S. L. P. Yasakethu, C. T. E. R. Hewage, W. A. C. Fernando, and A. M. Kondoz, "Quality analysis for 3D video using 2D video quality models," *IEEE Trans. Consum. Electron.*, vol. 54, no. 4, pp. 1969–1976, Nov. 2008.

[41] C. T. E. R. Hewage, S. T. Worrall, S. Dogan, and A. M. Kondoz, "Prediction of stereoscopic video quality using objective quality models of 2-D video," *Electron. Lett.*, vol. 44, no. 16, pp. 963–965, Jul. 2008.

[42] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Trans. Broadcast.*, vol. 50, no. 3, pp. 312–322, Sep. 2004.

[43] C. T. E. R. Hewage, S. T. Worrall, S. Dogan, S. Villette, and A. M. Kondoz, "Quality evaluation of color plus depth map-based stereoscopic video," *IEEE J. Sel. Topics Signal Process.*, vol. 3, no. 2, pp. 304–318, Apr. 2009.

[44] H. Malekmohamadi, A. Fernando, and A. Kondoz, *A New Reduced Reference Metric for Color Plus Depth 3D Video*. New York, NY, USA: Academic, 2014.

[45] T. Ebrahimi, L. Xing, J. You, and A. Perkis, "Assessment of stereoscopic crosstalk perception," *IEEE Trans. Multimedia*, vol. 14, no. 2, pp. 326–337, Apr. 2012.

[46] C. Galkandage, J. Calic, S. Dogan, and J.-Y. Guillemaut, "Stereoscopic video quality assessment using binocular energy," *IEEE J. Sel. Topics Signal Process.*, vol. 11, no. 1, pp. 102–112, Feb. 2017.

[47] J. Kim, H. Zeng, D. Ghadiyaram, S. Lee, L. Zhang, and A. C. Bovik, "Deep convolutional neural models for picture-quality prediction: Challenges and solutions to data-driven image quality assessment," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 130–141, Nov. 2017.

[48] L. Kang, P. Ye, Y. Li, and D. Doermann, "Convolutional neural networks for no-reference image quality assessment," in *Proc. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 1733–1740.

[49] W. Zhang, C. Qu, L. Ma, J. Guan, and R. Huang, "Learning structure of stereoscopic image for no-reference quality assessment with convolutional neural network," *Pattern Recognit.*, vol. 59, pp. 176–187, Nov. 2016.

[50] C. Fan, Y. Zhang, L. Feng, and Q. Jiang, "No reference image quality assessment based on multi-expert convolutional neural networks," *IEEE Access*, vol. 6, pp. 8934–8943, 2018.

[51] S. Zha, F. Luisier, W. Andrews, N. Srivastava, and R. Salakhutdinov, "Exploiting image-trained CNN architectures for unconstrained video classification," in *Proc. Brit. Mach. Vis. Conf. (BMVC)*, Sep. 2015, pp. 60.1–60.13.

[52] T. Pfister, K. Simonyan, J. Charles, and A. Zisserman, "Deep convolutional neural networks for efficient pose estimation in gesture videos," in *Proc. Asian Conf. Comput. Vis.*, 2015, pp. 538–552.

[53] Y. Li *et al.*, "No-reference video quality assessment with 3D shearlet transform and convolutional neural networks," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 26, no. 6, pp. 1044–1057, Jun. 2016.

[54] S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 221–231, Jan. 2013.

[55] A. Diba, A. M. Pazandeh, and L. Van Gool. (2016). "Efficient two-stream motion and appearance 3D CNNs for video classification." [Online]. Available: https://arxiv.org/abs/1608.08851

[56] P. Seuntiens, L. Meesters, and W. Ijsselsteijn, "Perceived quality of compressed stereoscopic images: Effects of symmetric and asymmetric JPEG coding and camera separation," *ACM Trans. Appl. Perception*, vol. 3, no. 2, pp. 95–109, Apr. 2006.

[57] *Methodology for the Subjective Assessment of the Quality of Television Pictures*, document ITU-R BT.500, 2002. Accessed: 2015. [Online]. Available: http://www.itu.int/rec/recommendation

[58] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Trans. Image Process.*, vol. 15, no. 2, pp. 430–444, Feb. 2006.

[59] L. Ma, X. Wang, Q. Liu, and K. N. Ngan, "Reorganized DCT-based image representation for reduced reference stereoscopic image quality assessment," *Neurocomputing*, vol. 215, pp. 21–31, Nov. 2016.

[60] J. Yang, Y. Lin, Z. Gao, Z. Lv, W. Wei, and H. Song, "Quality index for stereoscopic images by separately evaluating adding and subtracting," *PLoS ONE*, vol. 10, no. 12, p. e0145800, 2015.

[61] J. Yang, Y. Liu, Z. Gao, R. Chu, and Z. Song, "A perceptual stereoscopic image quality assessment model accounting for binocular combination behavior," *J. Vis. Commun. Image Represent.*, vol. 31, pp. 138–145, Aug. 2015.

[62] J. Yang *et al.*, "Quality assessment metric of stereo images considering cyclopean integration and visual saliency," *Inf. Sci.*, vol. 373, pp. 251–268, Dec. 2016.

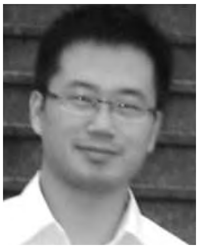
[63] N. Ozbek and A. M. Tekalp, "Unequal inter-view rate allocation using scalable stereo video coding and an objective stereo video quality measure," in *Proc. IEEE Int. Conf. Multimedia Expo*, Jun./Apr. 2008, pp. 1113–1116.

JIACHEN YANG (M'13) received the M.S. and Ph.D. degrees in communication and information engineering from Tianjin University, China, in 2005 and 2009, respectively. He was a Visiting Scholar with the Department of Computer Science, School of Science, Loughborough University, U.K. He is currently a Professor with Tianjin University. 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, where he is currently pursuing the M.S. degree with the School of Electrical and Information Engineering. His research interests include virtual reality and multimedia quality evaluation.

BIN JIANG received the B.S. and M.S. degrees in communication and information engineering from Tianjin University, Tianjin, China, in 2013 and 2016, where he is currently pursuing the Ph.D. degree with the School of Electrical and Information Engineering. He is also a Visiting Scholar with Ember-Riddle Aeronautical University, Daytona Beach, FL, USA. His research interests lie in computer vision, including stereo image, image quality, pattern recognition, and cyber-physical image processing.

HOUBING SONG (M'12–SM'14) received the Ph.D. degree in electrical engineering from the University of Virginia, Charlottesville, VA, USA, in 2012. He was a Faculty Member with West Virginia University from 2012 to 2017. In 2017, he joined the Department of Electrical, Computer, Software, and Systems Engineering, Embry-Riddle Aeronautical University, Daytona Beach, FL, USA, where he is currently an Assistant Professor and the Director of the Security and Optimization for Networked Globe Laboratory. He has edited four books and authored over 100 articles. His research interests include cyber-physical systems, internet of things, cloud computing, big data analytics, connected vehicle, wireless communications and networking, and optical communications and networking. He is a Senior Member of the ACM. He has been serving as an Associate Technical Editor for the *IEEE Communications Magazine* since 2017.



WEN LU (M'13) received the B.S. degree from the Xi'an University of Architecture and Technology in 1996 and the M.S. from the Beijing Institute of Technology in 2005. He was with the National Key Laboratory of Science and Technology on Aerospace Intelligence Control, Beijing, China. His research interests include pattern recognition, image classification, and retrieval.