Multimedia Systems (2024) 30:89
https://doi.org/10.1007/s00530-024-01285-0

REGULAR PAPER

360° video quality assessment based on saliency-guided viewport extraction

Fanxi Yang[1] · Chao Yang[1] · Ping An[1] · Xinpeng Huang[1]

Received: 6 October 2023 / Accepted: 8 February 2024 / Published online: 21 March 2024

Abstract
Due to the distortion of projection generated during the production of 360° video, most quality assessment algorithms used for 2D video have the problem of performance degradation. In this paper, we propose a full-reference 360° video quality assessment method, utilizing saliency to guide viewport extraction to eliminate the projection distortion. To be more specific, we first predict the visual saliency of each frame with a 360° saliency prediction network and then select the viewport that optimally represents the video frame through the optimal viewport positioning module (OVPM). Furthermore, we propose the attention-based three-dimensional convolutional neural network (3D CNN) quality assessment network to evaluate the video quality, in which 3D CNN convolution and attention modules can better capture the quality degradation of distorted viewports. Experimental results show that our method achieves superior performance in 360° video quality assessment tasks.

Keywords Video quality assessment · 360° video · Viewport selection · Saliency prediction

1 Introduction

Recently, with the rapid development of Internet technology, all kinds of media formats fill people's daily lives, and the sharing and transmission of images and videos are more frequent. There is also an increasing demand for high-quality images and videos. In the process of compression and uploading images and videos, the images and videos generated by non-professional devices are prone to damage, resulting in quality degradation, such as packet loss, blur, and Gaussian noise. Video quality assessment (VQA) has wide applications in many fields, such as image compression, video codec, video surveillance, and other necessary technologies.

As network bandwidth and display technology continue to advance, 360° videos have emerged as a novel and increasingly popular media format that is quickly gaining traction among the masses. The biggest difference between 360° video and traditional 2D video mainly lies in the process of stitching and projection. Stitching refers to the stitching of images shot by cameras from different angles onto a 360° sphere, while projection refers to the mapping of video from a sphere onto a plane to meet the needs of encoding, decoding, and transmission. Due to the influence of different cameras, the final video will often bring uneven lighting, ghosts, and other distortion factors in the stitching process. In the process of projection, pixels tend to be distorted and deformed to project the spherical content to the planar content, resulting in loss of information and distortion of the image [1].

Due to the influence of omnidirectional information and projection, common 360° videos often contain huge contents, and deformation occurs to different degrees with different image positions. Therefore, conventional feature extraction methods cannot adapt to the distortion deformation caused by projection, and the huge computational amount and complexity caused by stitching can hardly adapt to ordinary video quality assessment algorithms. Effectively

Communicated by Q. Shen.

✉ Chao Yang
yangchaoie@shu.edu.cn

Fanxi Yang
yangfx042702@shu.edu.cn

Ping An
anping@shu.edu.cn

Xinpeng Huang
huangxinpeng@shu.edu.cn

[1] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China

quantifying the quality degradation of 360° video and evaluating the quality of 360° video has great significance and practical application value for the research of 360° video processing.

To solve the above-mentioned problems, this paper proposes a quality assessment framework based on 360° saliency-guided viewport extraction, in which an innovative new algorithm based on saliency is designed to extract the most representative viewport. Combined with an outstanding saliency prediction algorithm [2], excellent results are achieved in the 360° video quality assessment task. The main contributions can be summarized as follows:

- To simulate the behavior of human eyes to perceive the quality of the region of interest when watching 360° videos, panoramic convolution was adopted to predict the saliency of video frames. We designed an algorithm to extract the viewport content to optimally represent the video content for quality assessment.
- We have designed a quality assessment network based on the attention mechanism. The network mainly uses a three-dimensional convolutional neural network (3D CNN) network to learn the spatiotemporal features of the video. Spatial and channel attention modules were designed to improve the network's ability to learn information, making it lightweight.
- The quality assessment network based on the attention mechanism proposed in this paper not only performs well in 360° video quality assessment tasks but also in 2D video quality assessment tasks.

2 Related works

2.1 VQA for 2D video

For 2D video, some quality assessment algorithms based on traditional hand-crafted feature extraction have achieved excellent results in video tasks for user-generated content. Considering the temporal effects of the human visual system, some image quality assessment (IQA) methods can be modified for VQA methods [3–7]. These methods employ different strategies and techniques to capture distortions and variations in dynamic video content. Some methods utilize motion compensation [8], while others address the perceptual effects of motion artifacts and quantify distortions through optical flow statistics [9]. AFViQ [10] leverages perceptual visual mechanisms in video quality assessment by introducing an enhanced foveal imaging model for generating the perceived video representation. He et al. [11] utilize 3D discrete cosine transform (DCT) to analyze and exploit energy and frequency distribution. Tu et al. [12] proposed the VIDEVAL algorithm to construct an initial feature set by selecting features from existing best-performing VQA models, which use machine learning models to learn important features. ChipQA [13] and the approach proposed by Wu et al. [14] both used the method of tracking motion trajectory to extract the features of video quality degradation. VMAF [15] is a video quality assessment tool launched by Netflix, designed to address situations where traditional indicators fail to reflect multiple scenes and features in videos.

As deep learning methods become more mature, many algorithms based on deep learning perform well in quality assessment tasks [16–18]. VSFA [19] leverages a pre-trained network for content dependency modeling and a recurrent network for temporal memory effect modeling. TLVQM [20] extracts low-complexity features from full video sequences and high-complexity features from representative frames for perceptual quality modeling. Inspired by PaQ-2-PiQ (from patch to picture) algorithm, Ying et al. [21] proposed the patchVQ algorithm. Based on the relationship between video patches and video clips, 3D convolution is adopted to extract spatiotemporal features from the spatiotemporal pool layer for learning. The RIR-net [22] makes creative use of a recurrent neural network to carry out quality assessment tasks. The framework designed by this network can better match the perception of distorted video quality in the human visual system (HVS). Most methods based on deep learning usually regard videos as static images and apply the pre-trained 2D CNN model to perform related tasks of video quality assessment on images. However, this strategy performs poorly in terms of motion sensitivity, as motion information is simply ignored. Because video quality is highly correlated with motion between successive frames, the full reference method proposed by Xu et al. [23], which applies 3D CNN to learn spatiotemporal features, had made progress in dealing with the compression artifacts and spatiotemporal continuity of video frames.

2.2 Quality assessment for 360° video/image

VQA for 360°. The existing 360° video quality assessment methods are generally divided into two categories: the traditional method and the learning-based method. The traditional method is an improvement on the 2D quality assessment algorithm. For example, the weighted to spherically uniform PSNR (WS-PSNR) [24] and the area-weighted spherical PSNR (AW-SPSNR) [25] both use weight distribution to balance the non-uniform sampling density when calculating the PSNR so that the non-uniformity of the spherical content mapped to the plane can be taken into account. Yu et al. [26] proposed S-PSNR resample the original frame with a set of uniformly distributed points on the sphere. Zakharchenko et al. [27] proposed a quality assessment based on the Craster parabolic projection (CPP), named CPP-PSNR. Xu et al. [28] introduced a

content-based perceptual PSNR (CP-PSNR) approach that calculates a weighted PSNR on the original 360° frame. Gao et al. [29] proposed a method that incorporates spatial and temporal considerations, evaluating distortions at the eye fixation level and providing an effective solution for integrating existing spatial video quality assessment metrics. All of these methods mentioned above obtain quality scores by calculating the PSNR of 360° video frame, but they do not achieve high accuracy. Yang et al. [30] proposed BP-QAVR, which uses a region of interest (ROI) map to calculate multi-level quality factors. Jiang et al. [31] proposed TB-VMAF, which utilizes elliptical projection, inverse projection, and bilinear interpolation to transform planar tiles into a sphere. By incorporating user head and eye movements to generate a tiled weighted map, they optimized the Video Multimethod Assessment Fusion (VMAF) method, achieving excellent results.

On the other hand, the algorithm based on deep learning has achieved great success in the field of 360° video quality assessment. Li et al. [32, 33] proposed the method of viewport-based convolutional neural network (V-CNN) for the full-reference 360° VQA task. In the V-CNN method, there are two stages, namely VP-net (Stage I) for extracting the potential viewport and VQ-net (Stage II) for calculating VQA. The proposed network also has two auxiliary tasks, the prediction of the potential viewport and the prediction of viewport significance, both of which have achieved good performance. The researchers also achieved excellent results with the no-referenced VQA algorithm. Meng et al. [34] proposed a method based on the fact that users show very consistent saliency preferences when consuming 360° content, the method was designed to combine the quality of highlighting viewport and the quality of quick scanning area of 360° video. In NR-OVQA [35], the 360° video is first projected by cube map projection (CMP) onto six equal-area 2D videos that are treated as inputs to CNN. Then, two-stream CNN models were built, and spatial and temporal quality features were modeled and learned.

IQA for 360°. In recent years, numerous IQA methods for 360° images have emerged based on deep learning. Many approaches have utilized visual and positional features of 360° images for quality prediction, yielding promising results [36–38]. Sun et al. [39] propose MC360IQA methods, which project each 360-degree image into six viewport images and use a multi-channel feature extraction model to learn the viewport feature expression. Xu et al. [40] proposed a Viewport-oriented graph convolution network (VGCN) that consists of two branches. One branch utilizes a viewport to calculate local quality scores, while the other branch uses DB-CNN for global quality score detection. It has achieved remarkable performance.

2.3 Saliency models on 360° video/image

Saliency Models on 360° **Image.** In recent years, researchers focused on the saliency prediction of 360° images. In [41, 42], researchers analyzed participants' gaze behavior with eye-tracking experiments in 360° videos. They improved the saliency model by incorporating specific gaze biases to adapt to this type of video and adjusted the weights of head movement data and eye movement data in their methods.

Saliency Models on 360° **Video.** To predict the behavior of participants in head-mounted displays (HMD), many researchers proposed methods for predicting eye movements (EM) and head movements (HM), which greatly aided in predicting saliency in 360° content [43, 44]. Xu et al. [45] proposed a deep reinforcement learning method that enables to predict the area that viewers are most likely to focus on. Additionally, there are various methods [46, 47] that utilize deep neural networks (DNNs) to predict the scanning paths of head and eye movements. Martin et al. [2] proposed a method that uses the panoramic convolutional network to predict 360° video saliency.

3 Methodology

We designed a quality assessment framework based on saliency prediction and viewport extraction. The framework of our proposed method is shown in Fig. 1, which can be divided into three parts: Video frame saliency prediction module, Optimal viewport positioning module, and Quality assessment network.

3.1 Frame saliency-guided viewport extraction

Previous studies have shown that image quality is highly related to visual saliency [48], and users exhibit consistent saliency preferences when consuming 360° content [49]. Therefore, the video frame saliency prediction algorithm adopted in this paper selects the network proposed by [2]. This approach introduces a panoramic convolutional network that learns feature relationships from a simple, non-distorted space, which is inspired by the approach proposed in [50]. Each point p on the sphere with latitude $\phi \in [\frac{-\pi}{2}, \frac{\pi}{2}]$ and longitude $\theta \in [-\pi, -\pi]$, there exists a tangent plane P located at p, whose coordinates $x, y \in P$ are related to a point on the sphere by its gnomonic projection. Using this network, it is no longer necessary to consider the distortion caused by the projection of panoramic video frames, so the global and local spatial dimension information is preserved.

The network uses a U-net-like structure composed of four encoder layers and four decoder layers. The encoder layer of the module consists of two panoramic blocks, and the decoder layer consists of three panoramic blocks. The

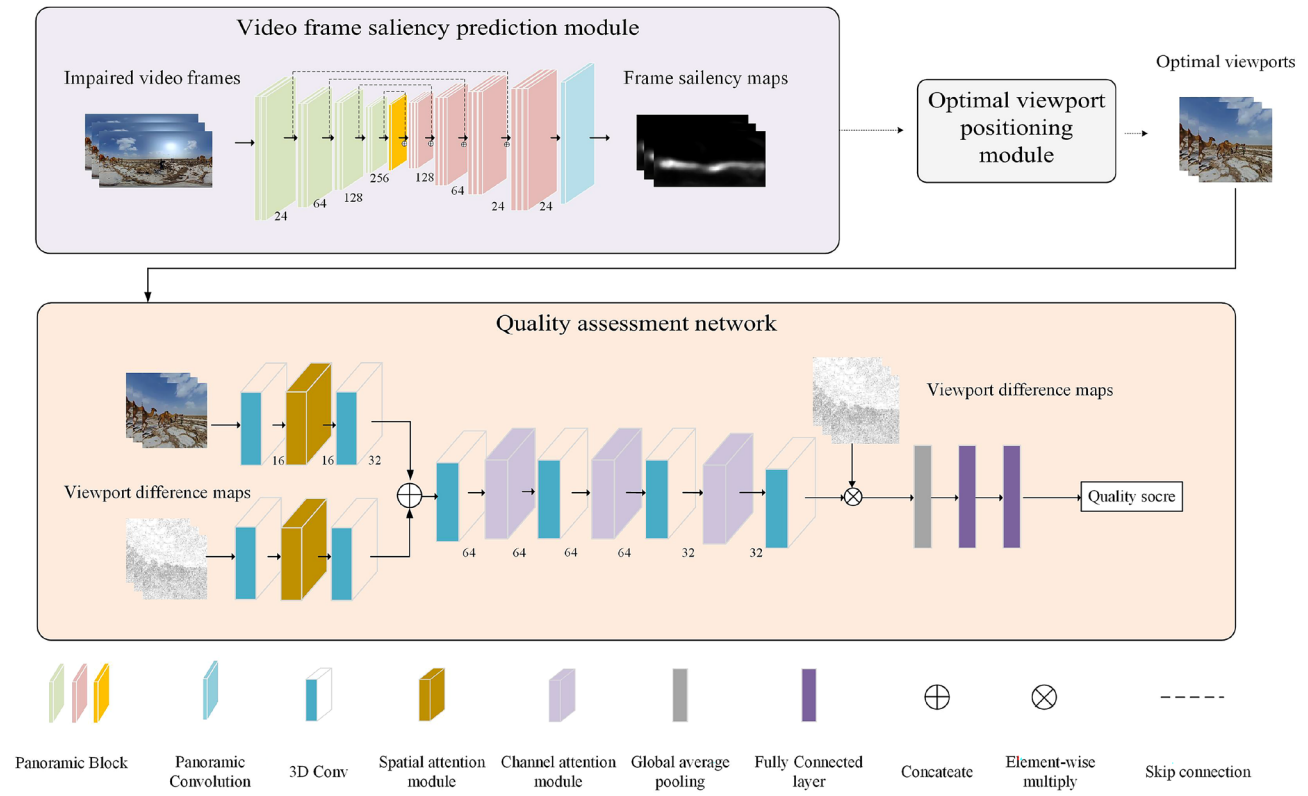

Fig. 1 The framework of our proposed method, the output channels of each convolution layer are denoted

structure of the panoramic block is illustrated in Fig. 2. To preserve the details of the final output image, the features of different resolutions on the encoder path are connected to the corresponding features on the decoder path through skipping connections. The model was trained using a publicly available dataset of panoramic image saliency [51]. This dataset successfully generates a saliency map from a single video frame and outperforms other advanced panoramic saliency prediction methods.

Due to the temporal and spatial continuity of video, unlike the content of 360° images, users cannot view multiple viewport contents on the same frame while watching 360° videos on HMD devices, making it difficult to evaluate the video quality. In continuous 360° videos, although users can view in all directions in HMD devices by rotating their heads, the human eyes can capture quality degradation at most on one viewport content in a certain frame. Figure 3 illustrates the distortion in the same region of equal rectangle projection (ERP) format and viewport. The ERP format causes a certain degree of deformation in the distorted area due to projection distortion, whereas the viewport is more representative of the content viewed in HMD. Therefore, using the viewport image for distortion evaluation is more appropriate. Therefore, to simulate the process of the human eye evaluating video quality, designing the correct viewport selection method becomes crucial.

Inspired by this behavior, the optimal viewport positioning module (OVPM) is introduced to extract the viewport that optimally represents the video frame. For each 360° video, we sample $n(0 < n \leq N-2)$ video clips at intervals, where N is the total number of frames. Each clip consists of three frames, center frame I_t and its adjacent frames $\{I_{t-1}, I_{t+1}\}$. A total of $3 \times n$ frames are sent into the module for saliency prediction which generates the video frame saliency map of ERP format. We set the horizontal and vertical field of view (FOV) angle of each viewport to 90° as shown in Fig. 4, and extraction of viewports is carried out by referring to CMP and gnomonic projection. To extract the viewport, the viewport center point

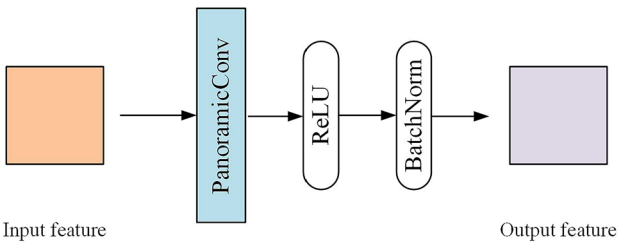

Fig. 2 The structure of the Panoramic Block

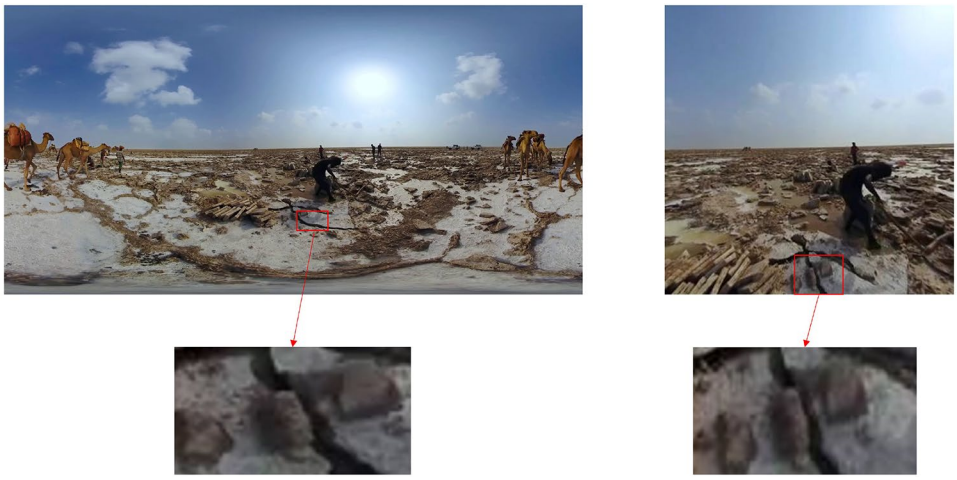

Fig. 3 The distortion difference between ERP format and viewport

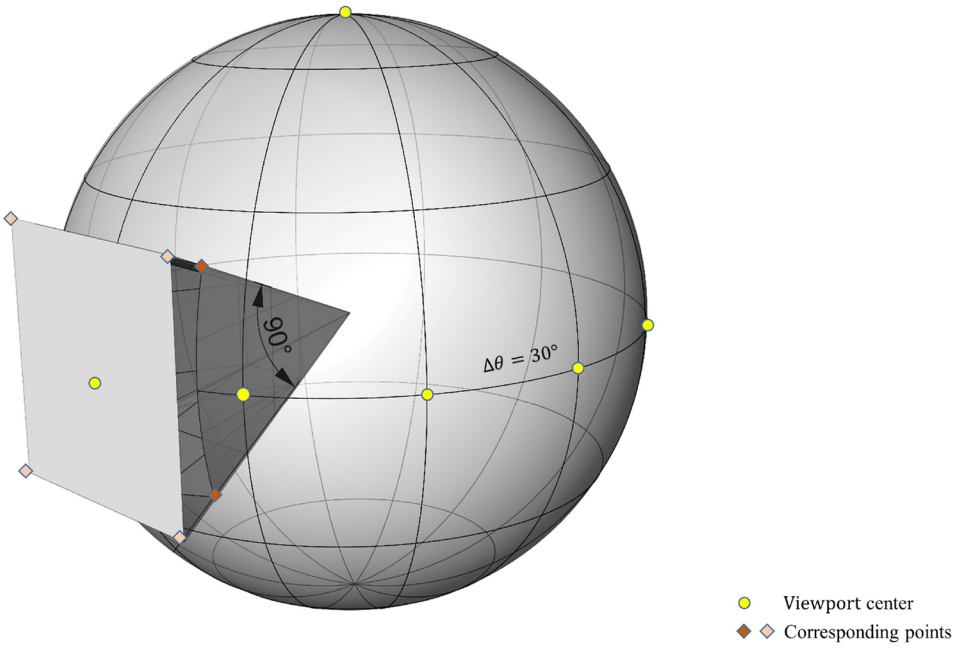

Fig. 4 A diagram of the viewport generated using the gnomonic projection, the example of 14 viewport centers on the sphere, and corresponding points on sphere surface and tangent plane

$p^0_{\mathrm{ERP}} = (x^0, y^0)$ of saliency image in ERP format is projected back to the sphere $p^0_s = (\phi^0, \theta^0)$ since the angle of FOV is fixed when the user is viewing 360° content, each viewport can be represented as a projection of the tangent plane centered on $p^0 = (\phi^0, \theta^0)$. Initially, the saliency image is transformed into an ERP projection format, which is subsequently mapped back onto a sphere. The conversion formula between the ERP domain and the sphere domain is presented as follows:

$$\begin{cases} \phi = \frac{2\pi u}{W} - \pi \\ \theta = \frac{-\pi v}{H} + \frac{\pi}{2} \end{cases} \tag{1}$$

where H and W are the height and width of ERP frames, u and v are the coordinates of the ERP domain. $\phi \in [\frac{-\pi}{2}, \frac{\pi}{2}]$ and $\theta \in [-\pi, -\pi]$ are latitude and longitude on the sphere.

Based on the gnomonic projection and the angle of FOV, the coordinates of the sphere domain are projected onto the tangent plane of the viewport's central point to determine the corresponding viewport. The conversion formula between the tangent plane and the sphere domain is presented as follows:

$$\begin{aligned} x(\phi, \theta) &= \frac{\cos\phi \sin\left(\theta - \theta_{\Pi_0}\right)}{\sin\phi_{\Pi_0} \sin\phi + \cos\phi_{\Pi_0} \cos\phi \cos\left(\theta - \theta_{\Pi_0}\right)} \\ y(\phi, \theta) &= \frac{\cos\phi_{\Pi_0} \sin\phi - \sin\phi_{\Pi_0} \cos\phi \cos\left(\theta - \theta_{\Pi_0}\right)}{\sin\phi_{\Pi_0} \sin\phi + \cos\phi_{\Pi_0} \cos\phi \cos\left(\theta - \theta_{\Pi_0}\right)} \end{aligned} \tag{2}$$

where (ϕ, θ) represents the coordinates of the sphere, (x, y) represents the coordinates of the tangent plane, $(\phi_{\Pi_0}, \theta_{\Pi_0})$

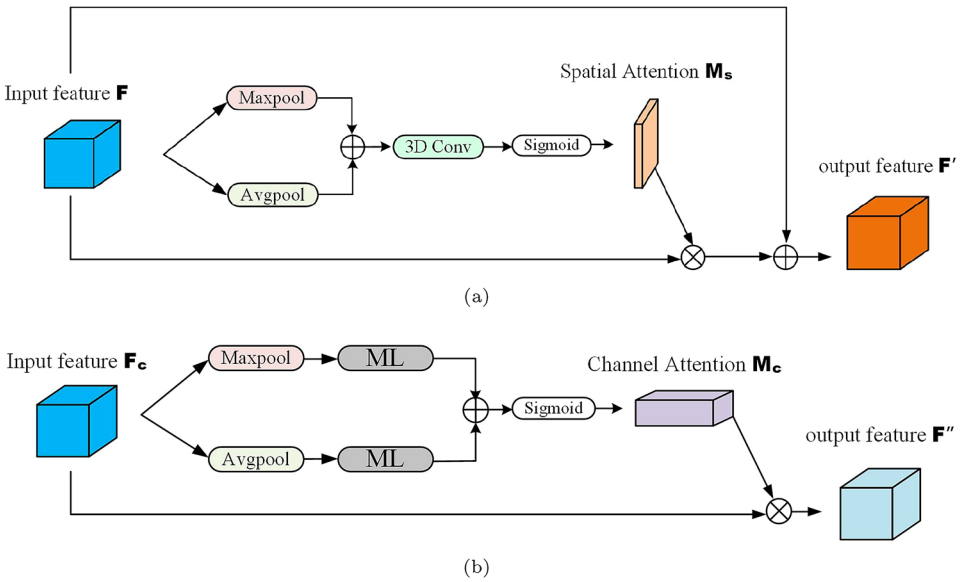

Fig. 5 Diagram of attention modules: **a** is the architecture of the spatial attention module, **b** represents the illustration of the channel attention model

represents the center point of the viewport on the sphere domain.

In this module, as shown in Fig. 4, 14 alternative potential viewport centers $\{p_1, p_2, p_3, \ldots, p_{14}\}$ are set, of which 12 are located on the equator of the saliency projection sphere, and the other two are located on the poles of the sphere. The centers of the adjacent viewport on the equator $\Delta\theta = 30°$. The 14 alternative viewports fully covered all the information on the video frame. The maximum value of the saliency values on the 14 viewports was calculated $Max\{S_1, S_2, S_3, \ldots, S_{14}\}$, where the saliency value S_i was the sum of pixel values in the gray-scale saliency map. The viewport with the highest saliency value represents the content that the human eye is most likely to watch in this frame, which is consistent with the area of interest of the human eye watching the video. The corresponding viewport content is then input into the quality assessment network as the representative of the video frame for quality assessment.

3.2 Quality assessment network

As the viewport content is presented in the HMD for virtual reality applications with almost no impact of projection distortion, we input the optimal viewport content extracted from the previous module to the quality assessment network. Inspired by the 2D full-reference video quality assessment algorithm based on Just Noticeable Difference (JND) [23], the structure of the quality assessment network is shown in Fig. 1. Before inputting viewport content into the network, we calculate the viewport difference map as follows:

$$V_t^{\text{diff}} = \left| \frac{2\ln(255) - \ln\left((V_t^{\text{ref}} - V_t^{\text{dist}})^2 + 1\right)}{2\ln 255} \right| \tag{3}$$

where V_t^{ref} is the viewport content extracted from the reference 360° video frame. For each pixel in both impaired viewport content V_t^{dist} and reference viewport content V_t^{ref}, the alignment is implemented by bilinear interpolation on the frame at the corresponding location. The residual values are normalized to ensure that their range is between 0 and 1, making negative pixel values impossible. Then, both the impaired viewport V_t^{dist} and the difference map V_t^{diff} are input into the quality assessment network. We use two 3D convolution layers to downsample the viewport content for spatiotemporal feature learning. The spatial attention (SA) module is introduced to give weight to the selection of spatial features, to optimize the network's learning ability for information. The illustration of the spatial attention module given in Fig. 5a can be summarized as:

$$F' = (1 + M_s(F)) \otimes F \tag{4}$$

$$M_s(F) = \sigma(f^{7\times7}(\text{AvgPool}(F) \oplus \text{MaxPool}(F))) \tag{5}$$

where F is the feature map after two 3D convolution layers, σ denotes the sigmoid function, $f^{7\times7}$ represents a convolution operation with the filter size of 7×7, $\otimes$ and $\oplus$ means element-wise multiplication and concatenation.

After channel concatenation of the two feature maps F', four 3D convolution layers and corresponding channel attention (CA) modules are introduced to further select the features. The illustration of the channel attention shown in Fig. 5b can be obtained as follows:

$$F'' = M_c(F_c) \otimes F_c \tag{6}$$

$$M_s(F_c) = \sigma(ML(\text{AvgPool}(F_c)) + \text{ML}(\text{MaxPool}(F_c))) \tag{7}$$

where the ML denotes multi-layers, which consist of two convolution layers and a ReLU layer.

Before sending the final feature map into the global average pooling layer, we multiplied the difference map V_t^{diff} by the feature map to represent the degree of perceived distortions. Next, two fully connected layers are used to conduct nonlinear mapping between distortions and subjective scores. The network's training loss can be obtained as:

$$\mathcal{L} = \frac{1}{K}\left\| f_\delta(\mathrm{x_n}) - y_n \right\|_2^2 + \lambda L_2 \tag{8}$$

where K denotes the total number of impaired videos in the training set, $\mathrm{x_n}$ and y_n denote the n-th input video pair and the subjective score of the video, respectively. δ is the parameters of the whole network that need to be trained. λ and L_2 represent a hyper-parameter and a regularization term, respectively.

4 Experimental results

In this section, we evaluate the proposed method on the commonly used 360° VQA database VQA-ODV [15] and 2D VQA database CSIQ-VQA [52]. The Spearman rank order correlation coefficient (SROCC), Pearson linear correlation coefficient (PLCC), Kendall rank order correlation coefficient (KROCC), and root mean square error (RMSE) are used as evaluation criteria. Our method was compared with several state-of-the-art methods for VQA on 360° video and 2D video. We designed experiments to verify the number of viewport content from one frame can be used as network input to achieve optimal performance. The effectiveness of the optimal viewport positioning module was discussed by replacing the optimal viewport position with a fixed viewport position, and we also studied the effect of the FOV angle. We conducted experiments to demonstrate the effectiveness of attention modules and the impact of the number of frames extracted from a video on model performance.

4.1 Datasets and training details

VQA-ODV dataset. This dataset consists of 60 reference videos and 540 distorted videos generated from 3 project patterns, i.e., ERP, reshaped cube map projection (RCMP) and truncated square pyramid projection (TSP), each of which contained three compression levels generated using H.256, i.e., QP = 27, 37 and 42. The resolution of the dataset contains 4K (3840 × 1920) to 8K (7680 × 3840).

CSIQ-VQA dataset. This dataset consists of 12 reference videos and 216 distorted videos generated from 6 distortion types, i.e., H.264/AVC compression, H.264 video with packet loss rate, MJPEG compression, Wavelet compression, White noise, and HEVC compression, each of which also contained three compression levels. The videos in this dataset all have a resolution of 832 × 480.

In our experiments, we randomly select 80% of the reference videos for training, and the remaining 20% are used for testing. Once a reference video is divided into the training or testing set, all distorted videos generated from it will be put into the same set. We use the Adam optimizer to back-propagate gradients and regularization, the initial learning rate for the VQA-ODV dataset and CSIQ-VQA dataset is 7e−4 and 3e−4, respectively, and the weight decay is set 5e−3 and 3e−3, respectively. The learning rate would multiply by 0.9 if the loss saturates for 5 epochs. The number of video clips n put into the method is 9 and 11, respectively. Our experiments are implemented based on the PyTorch framework and run on an NVIDIA GTX3060TI GPU with 8 G memory, an i7-11700k, and 32 GigaBytes of RAM.

4.2 Performance on 360° videos

We compare the performance of our proposed method with some state-of-the-art methods for VQA on VQA-ODV dataset [33] as follows, WS-PSNR [24], S-PSNR [26], CPP-PSNR [27], NR-OVQA [35], MC360IQA [39], VGCN [40], BP-QAVR [30], OV-PSNR [29], TB-VMAF [31], V-CNN [32, 33]. Among them, WS-PSNR, S-PSNR, and CPP-PSNR are PSNR-based methods for 360° VQA, which are implemented using the code[1]; MC360IQA and VGCN are no-reference methods for 360° image, which achieve the excellent result on 360° IQA database. NR-OVQA is a no-reference method for 360° VQA; BP-QAVR, OV-PSNR, and TB-VMAF are the traditional full-reference methods for 360° VQA; V-CNN is the full-reference deep learning-based method.

The results are shown in Table 1. It can be observed that V-CNN achieves the best performance in PLCC, SROCC, and RMSE, and our proposed method is superior to any other methods in KROCC, only inferior to V-CNN in these three indexes and reaches the same order of magnitude. Moreover, in terms of model complexity, our model parameters and FLOPs are far less than those of other deep learning-based methods. This implies that our proposed method requires less computation to calculate the loss and is better at preventing overfitting.

4.3 Performance on 2D videos

We compare the performance of our proposed quality assessment network with some state-of-the-art methods for 2D VQA on CSIQ-VQA dataset [52] as follows, ChipQA [13], Wu et al. [14], VMAF [15], RIR-net [22], C3DVQA [23].

[1] Available: https://github.com/Samsung/360tools.

Table 1 Performance comparison on VQA-ODV dataset with competing methods, the best two results are in bold

Methods	PLCC↑	SROCC↑	KROCC↑	RMSE↓	Parameter	FLOPs
S-PSNR [26]	0.5787	0.6077	0.4571	14.0340	–	–
WS-PSNR [24]	0.5364	0.5745	0.4158	14.9314	–	–
CPP-PSNR [27]	0.5627	0.6041	0.4444	14.5020	–	–
NR-OVQA [17]	0.7598	0.7972	0.6286	7.7006	43.10M	418.77G
MC360IQA [39]	0.7589	0.7831	0.6075	7.7134	22.40M	22.17G
VGCN [40]	0.8032	0.8122	0.6144	7.0562	26.66M	220.00G
BP-QAVR [30]	0.6588	0.6801	0.4780	8.9112	0.66M	2.46G
OV-PSNR [29]	0.7351	0.7314	0.5019	7.1002	–	–
TB-VMAF [31]	0.8863	0.8721	0.6952	5.3469	–	–
V-CNN [33]	**0.9196**	**0.9140**	**0.7432**	**4.6527**	5.71M	11.24G
Proposed	**0.9165**	**0.9068**	**0.7493**	**5.1349**	3.17M	4.36G

Table 2 Performance comparison on CSIQ-VQA dataset with competing methods, the best results are in bold

Methods	PLCC↑	SROCC↑
VMAF [15]	0.6581	0.6377
ChipQA [13]	0.5222	0.5336
Wu [14]	0.8850	0.9076
RIR-net [22]	0.8426	0.8574
C3DVQA [23]	0.9043	0.9152
proposed	**0.9137**	**0.9246**

Among them, VMAF uses several metrics to calculate and aggregate the frame quality score to obtain the final video quality score; ChipQA and Wu [14] are both the machine learning-based methods for 2D VQA; RIR-net and C3DVQA are the full-reference deep learning-based methods.

The results are shown in Table 2. It can be observed from the results that the quality assessment network with the attention mechanism proposed by us is optimal in both PLCC and SROCC metrics. It is proved that the attention mechanism can play a very positive role in the perception of quality degradation in 2D video quality assessment.

4.4 Ablation study

To study the performance of the proposed method with different viewport numbers in one single video frame, we conduct VQA experiments with 1, 3, and 5 viewports. These selected viewports are with the highest saliency values calculated by the OVPM module. We can achieve the best performance when the number of viewports is 1 through Table 3. This result is consistent with the user's behavior when watching 360° videos.

We validate the effectiveness of the OVPM. Inspired by the view direction [35] in CMP format. We replace the module by setting the fixed viewport center of each frame, including $p_1(0°\text{N}, 0°\text{E})$, $p_2(0°\text{N}, 90°\text{W})$, and $p_3(0°\text{N}, 90°\text{E})$

Table 3 Ablation study on the number of viewports extracted in one single video frame on the VQA-ODV dataset

Viewport Number	PLCC↑	SROCC↑	KROCC↑	RMSE↓
1	0.9165	0.9068	0.7492	5.1349
3	0.7818	0.7902	0.6159	9.1781
5	0.8217	0.8111	0.6317	7.0972

Table 4 Ablation study on the optimal viewport positioning module on the VQA-ODV dataset

Viewport center	PLCC↑	SROCC↑	KROCC↑	RMSE↓
$p_1(0°\text{N}, 0°\text{E})$	0.8419	0.8221	0.6444	7.1005
$p_2(0°\text{N}, 90°\text{W})$	0.7891	0.7640	0.5841	8.5235
$p_3(0°\text{N}, 90°\text{E})$	0.8511	0.8399	0.6444	7.2054
ours with OVPM	0.9165	0.9068	0.7492	5.1349

correspond to front, left, and right view in CMP format. The corresponding viewport contents are the input of the quality assessment network. Valuation results are shown in Table 4, as expected, the performance with the OVPM is significantly better than the performance with a fixed viewport center.

When human eyes watch 360° content in HMD devices, FOV will inevitably affect human eyes' perception of quality. Five FOV angles during viewport extraction were set for ablation experiments. The experiment results are shown in Table 6. The best performance is obtained when the angle of FOV is 90°, and it was proved through experiments that the FOV angle also had a great influence on the quality assessment of 360° video.

To validate the effectiveness of the proposed attention module, we conducted experiments to investigate the impact of the network without SA, the network without CA, and the network with no attention modules. The results of the ablation study, as shown in Table 5, indicate

Table 5 Ablation study on attention module on the VQA-ODV dataset

SA	CA	PLCC↑	SROCC↑	KROCC↑	RMSE↓
		0.8217	0.8142	0.5988	6.5247
✓		0.8483	0.8325	0.6251	6.0333
	✓	0.8617	0.8564	0.6471	5.8117
✓	✓	0.9165	0.9068	0.7492	5.1349

Table 6 Ablation study on the optimal FOV angle on the VQA-ODV dataset

FOV angle	PLCC↑	SROCC↑	KROCC↑	RMSE↓
70°	0.8028	0.7907	0.5905	8.5247
80°	0.8334	0.8085	0.6159	6.8117
90°	0.9165	0.9068	0.7492	5.1349
100°	0.8990	0.8837	0.6970	5.1752
110°	0.8496	0.8432	0.6762	6.5371

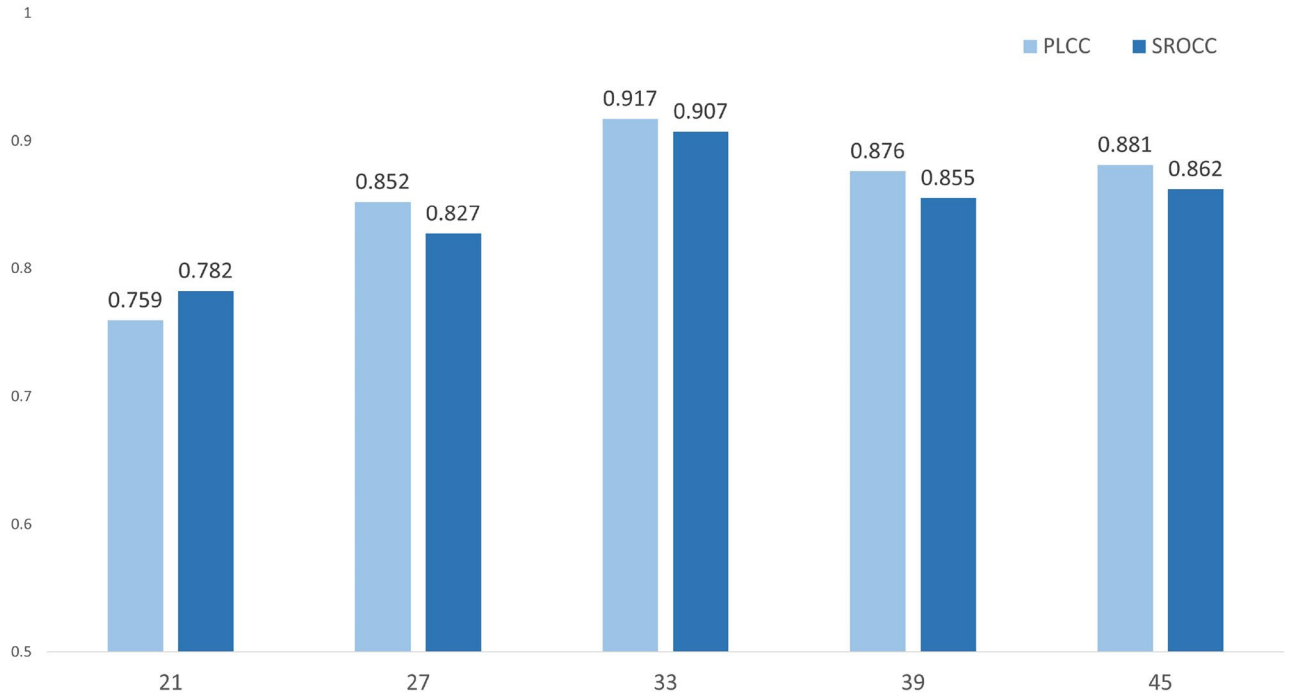

Fig. 6 SROCC and PLCC result of the proposed method trained with different numbers of frames per video

that both spatial and channel attention modules significantly improve the performance of the network.

Based on our empirical observations, we have found that the number of frames extracted from a video has a significant impact on both the model's performance and the computational cost. To investigate this impact, we conducted an ablation experiment where we repeated the experiment with each video being extracted at different numbers of frames. Specifically, we tested 21, 27, 33, 39, and 45 frames per video. The results of the ablation experiment, as shown in Fig. 6, indicate that setting the number of frames to 33 per video results in the best performance. Our experiments also revealed that increasing the number of frames does not necessarily lead to better performance; instead, it consumes a significant amount of computational resources and may result in overfitting. Therefore, we opted for an appropriate number of frames as input.

4.5 Performance influence of different FOV angles

When human eyes watch 360° content in HMD devices, FOV will inevitably affect human eyes' perception of quality. Five FOV angles were set during viewport extraction for ablation experiments. The experiment results are shown in Table 6. The best performance is obtained when the angle of FOV is 90°, and it was proven through experiments that the FOV angle also had a great influence on the quality assessment of 360° video.

5 Conclusion

To evaluate 360° video quality, we propose a deep learning method based on saliency prediction and viewport extraction. The saliency prediction network predicts video frame saliency, and the viewport that most attracts human attention is extracted based on saliency for quality assessment. The

experimental results show that our method achieves comparable performance with the state-of-the-art method with much fewer model parameters. The quality assessment network with attention mechanisms can also achieve excellent results in 2D video quality assessment tasks.

Acknowledgements This work was supported in part by the NSFC under Grant 62371279, 62171002, 61901252, 62071287, 62020106011, 62371278, and Science and Technology Commission of Shanghai Municipality under Grant 22ZR1424300.

Author contributions Fanxi Yang: contributed to the conception of the study, performed the experiment, and wrote the manuscript text Chao Yang: contributed significantly to the analysis and wrote the manuscript text Ping An: helped perform the analysis with constructive discussions, and reviewed the manuscript. Xinpeng Huang: helped perform the analysis with constructive discussions, and reviewed the manuscript.

Data availability The datasets generated and/or analyzed during the current study are available from the corresponding author on reasonable request.

Declarations

Conflict of interest The authors declare no competing interests.

References

1. Xu, M., Li, C., Zhang, S., Le Callet, P.: State-of-the-art in 360 video/image processing: perception, assessment and compression. IEEE J. Sel. Top. Signal Process. **14**(1), 5–26 (2020)
2. Martin, D., Serrano, A., Masia, B.: Panoramic convolutions for 360 single-image saliency prediction. In: CVPR workshop on computer vision for augmented and virtual reality, vol. 2 (2020)
3. Seshadrinathan, K., Bovik, A.C.: Temporal hysteresis model of time varying subjective video quality. In: 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP), pp. 1153–1156, IEEE (2011)
4. Wang, Y., Jiang, T., Ma, S., Gao, W.: Novel spatio-temporal structural information based video quality metric. IEEE Trans. Circuits Syst. Video Technol. **22**(7), 989–998 (2012)
5. Seshadrinathan, K., Bovik, A.C.: Motion tuned spatio-temporal quality assessment of natural videos. IEEE Trans. Image Process. **19**(2), 335–350 (2009)
6. Vu, P.V., Vu, C.T., Chandler, D.M.: A spatiotemporal most-apparent-distortion model for video quality assessment. In: 2011 18th IEEE international conference on image processing, pp. 2505–2508 (2011). IEEE
7. Larson, E.C., Chandler, D.M.: Most apparent distortion: full-reference image quality assessment and the role of strategy. J. Electron. Imaging **19**(1), 011006–011006 (2010)
8. Moorthy, A.K., Bovik, A.C.: Efficient video quality assessment along temporal trajectories. IEEE Trans. Circuits Syst. Video Technol. **20**(11), 1653–1658 (2010)
9. Manasa, K., Channappayya, S.S.: An optical flow-based full reference video quality assessment algorithm. IEEE Trans. Image Process. **25**(6), 2480–2492 (2016)
10. You, J., Ebrahimi, T., Perkis, A.: Attention driven foveated video quality assessment. IEEE Trans. Image Process. **23**(1), 200–213 (2013)
11. He, L., Lu, W., Jia, C., Hao, L.: Video quality assessment by compact representation of energy in 3d-dct domain. Neurocomputing **269**, 108–116 (2017)
12. Tu, Z., Wang, Y., Birkbeck, N., Adsumilli, B., Bovik, A.C.: Ugc-vqa: Benchmarking blind video quality assessment for user generated content. IEEE Trans. Image Process. **30**, 4449–4464 (2021)
13. Ebenezer, J.P., Shang, Z., Wu, Y., Wei, H., Sethuraman, S., Bovik, A.C.: Chipqa: No-reference video quality prediction via space-time chips. IEEE Trans. Image Process. **30**, 8059–8074 (2021)
14. Wu, J., Liu, Y., Dong, W., Shi, G., Lin, W.: Quality assessment for video with degradation along salient trajectories. IEEE Trans. Multimedia **21**(11), 2738–2749 (2019)
15. Rassool, R.: Vmaf reproducibility: Validating a perceptual practical video quality metric, pp. 1–2 (2017). IEEE
16. Li, Y., Po, L.-M., Cheung, C.-H., Xu, X., Feng, L., Yuan, F., Cheung, K.-W.: No-reference video quality assessment with 3d shearlet transform and convolutional neural networks. IEEE Trans. Circuits Syst. Video Technol. **26**(6), 1044–1057 (2015)
17. Liu, W., Duanmu, Z., Wang, Z.: End-to-end blind quality assessment of compressed videos using deep neural networks., pp. 546–554 (2018)
18. Zhang, Y., Gao, X., He, L., Lu, W., He, R.: Blind video quality assessment with weakly supervised learning and resampling strategy. IEEE Trans. Circuits Syst. Video Technol. **29**(8), 2244–2255 (2018)
19. Li, D., Jiang, T., Jiang, M.: Quality assessment of in-the-wild videos. In: Proceedings of the 27th ACM international conference on multimedia, pp. 2351–2359 (2019)
20. Korhonen, J.: Two-level approach for no-reference consumer video quality assessment. IEEE Trans. Image Process. **28**(12), 5923–5938 (2019)
21. Ying, Z., Mandal, M., Ghadiyaram, D., Bovik, A.: Patch-vq:'patching up'the video quality problem. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 14019–14029 (2021)
22. Chen, P., Li, L., Ma, L., Wu, J., Shi, G.: Rirnet: Recurrent-in-recurrent network for video quality assessment. In: Proceedings of the 28th ACM international conference on multimedia, pp. 834–842 (2020)
23. Xu, M., Chen, J., Wang, H., Liu, S., Li, G., Bai, Z.: C3dvqa: Full-reference video quality assessment with 3d convolutional neural network. In: ICASSP 2020-2020 IEEE International conference on acoustics, speech and signal processing (ICASSP), pp. 4447–4451. IEEE (2020)
24. Sun, Y., Lu, A., Yu, L.: Weighted-to-spherically-uniform quality evaluation for omnidirectional video. IEEE Signal Process. Lett. **24**(9), 1408–1412 (2017)
25. Xiu, X., He, Y., Ye, Y., Vishwanath, B.: An evaluation framework for 360-degree video compression. In: 2017 IEEE visual communications and image processing (VCIP), pp. 1–4 (2017). IEEE
26. Yu, M., Lakshman, H., Girod, B.: A framework to evaluate omnidirectional video coding schemes. In: 2015 IEEE international symposium on mixed and augmented reality, pp. 31–36 (2015). IEEE
27. Zakharchenko, V., Choi, K.P., Park, J.H.: Quality metric for spherical panoramic video. Opt. Photon. Inform. Process. X **9970**, 57–65 (2016)
28. Xu, M., Li, C., Chen, Z., Wang, Z., Guan, Z.: Assessing visual quality of omnidirectional videos. IEEE Trans. Circuits Syst. Video Technol. **29**(12), 3516–3530 (2018)
29. Gao, P., Zhang, P., Smolic, A.: Quality assessment for omnidirectional video: a spatio-temporal distortion modeling approach. IEEE Trans. Multimedia **24**, 1–16 (2020)
30. Yang, S., Zhao, J., Jiang, T., Wang, J., Rahim, T., Zhang, B., Xu, Z., Fei, Z.: An objective assessment method based on multi-level factors for panoramic videos. In: 2017 IEEE visual

communications and image processing (VCIP), pp. 1–4 (2017). IEEE
31. Jiang, Z., Xu, Y., Sun, J., Hwang, J.-N., Zhang, Y., Appleby, S.C.: Tile-based panoramic video quality assessment. IEEE Trans. Broadcast. **68**(2), 530–544 (2021)
32. Li, C., Xu, M., Jiang, L., Zhang, S., Tao, X.: Viewport proposal cnn for 360° video quality assessment. In: 2019 IEEE/CVF conference on computer vision and pattern recognition (CVPR), pp. 10169–10178 (2019). IEEE
33. Xu, M., Jiang, L., Li, C., Wang, Z., Tao, X.: Viewport-based CNN: a multi-task approach for assessing 360° video quality. IEEE Trans. Pattern Anal. Mach. Intell. **44**(4), 2198–2215 (2020)
34. Meng, Y., Ma, Z.: Viewport-based omnidirectional video quality assessment: database, modeling and inference. IEEE Trans. Circuits Syst. Video Technol. **32**(1), 120–134 (2021)
35. Chai, X., Shao, F.: Blind quality assessment of omnidirectional videos using spatio-temporal convolutional neural networks. Optik **226**, 165887 (2021)
36. Kim, H.G., Lim, H.-T., Ro, Y.M.: Deep virtual reality image quality assessment with human perception guider for omnidirectional image. IEEE Trans. Circuits Syst. Video Technol. **30**(4), 917–928 (2019)
37. Zhou, Y., Sun, Y., Li, L., Gu, K., Fang, Y.: Omnidirectional image quality assessment by distortion discrimination assisted multi-stream network. IEEE Trans. Circuits Syst. Video Technol. **32**(4), 1767–1777 (2021)
38. Chai, X., Shao, F., Jiang, Q., Meng, X., Ho, Y.-S.: Monocular and binocular interactions oriented deformable convolutional networks for blind quality assessment of stereoscopic omnidirectional images. IEEE Trans. Circuits Syst. Video Technol. **32**(6), 3407–3421 (2021)
39. Sun, W., Min, X., Zhai, G., Gu, K., Duan, H., Ma, S.: Mc360iqa: a multi-channel CNN for blind 360-degree image quality assessment. IEEE J. Sel. Top. Signal Process. **14**(1), 64–77 (2019)
40. Xu, J., Zhou, W., Chen, Z.: Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks. IEEE Trans. Circuits Syst. Video Technol. **31**(5), 1724–1737 (2020)
41. Rai, Y., Le Callet, P., Guillotel, P.: Which saliency weighting for omni directional image quality assessment? In: 2017 Ninth international conference on quality of multimedia experience (QoMEX), pp. 1–6 (2017). IEEE
42. Sitzmann, V., Serrano, A., Pavel, A., Agrawala, M., Gutierrez, D., Masia, B., Wetzstein, G.: Saliency in vr: How do people explore virtual environments? IEEE Trans. Visual Comput. Graphics **24**(4), 1633–1642 (2018)
43. Xu, Y., Dong, Y., Wu, J., Sun, Z., Shi, Z., Yu, J., Gao, S.: Gaze prediction in dynamic 360 immersive videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 5333–5342 (2018)
44. Cheng, H.-T., Chao, C.-H., Dong, J.-D., Wen, H.-K., Liu, T.-L., Sun, M.: Cube padding for weakly-supervised saliency prediction in 360 videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp. 1420–1429 (2018)
45. Xu, M., Song, Y., Wang, J., Qiao, M., Huo, L., Wang, Z.: Predicting head movement in panoramic video: a deep reinforcement learning approach. IEEE Trans. Pattern Anal. Mach. Intell. **41**(11), 2693–2708 (2018)
46. Li, F., Bai, H., Zhao, Y.: Visual attention guided eye movements for 360 degree images. In: 2017 Asia-Pacific signal and information processing association annual summit and conference (APSIPA ASC), pp. 506–511 (2017). IEEE
47. Assens Reina, M., Giro-i-Nieto, X., McGuinness, K., O'Connor, N.E.: Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In: Proceedings of the IEEE international conference on computer vision workshops, pp. 2331–2338 (2017)
48. Zhang, L., Shen, Y., Li, H.: Vsi: A visual saliency-induced index for perceptual image quality assessment. IEEE Trans. Image Process. **23**(10), 4270–4281 (2014)
49. Xu, M., Song, Y., Wang, J., Qiao, M., Huo, L., Wang, Z.: Predicting head movement in panoramic video: a deep reinforcement learning approach. IEEE Trans. Pattern Anal. Mach. Intell. **41**(11), 2693–2708 (2018)
50. Coors, B., Condurache, A.P., Geiger, A.: Spherenet: Learning spherical representations for detection and classification in omnidirectional images. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 518–533 (2018)
51. Monroy, R., Lutz, S., Chalasani, T., Smolic, A.: Salnet360: Saliency maps for omni-directional images with CNN. Sig. Process. Image Commun. **69**, 26–34 (2018)
52. Vu, P.V., Chandler, D.M.: Vis 3: an algorithm for video quality assessment via analysis of spatial and spatiotemporal slices. J. Electron. Imaging **23**(1), 013016–013016 (2014)