This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3044458, IEEE Transactions on Multimedia

Quality Assessment for Omnidirectional Video: A Spatio-Temporal Distortion Modeling Approach

Pan Gao, Pengwei Zhang, and Aljosa Smolic

***Abstract*—Omnidirectional video, also known as 360-degree video, has become increasingly popular nowadays due to its ability to provide immersive and interactive visual experiences. However, the ultra high resolution and the spherical observation space brought by the large spherical viewing range make omnidirectional video distinctly different from traditional 2D video. To date, the video quality assessment (VQA) for omnidirectional video is still an open issue. The existing VQA metrics for omnidirectional video only consider the spatial characteristics of distortions, but the temporal changes of spatial distortions can also considerably influence human visual perception. In this paper, we propose a spatiotemporal modeling approach to evaluate the quality of the omnidirectional video. Firstly, we construct a spatiotemporal quality assessment unit to evaluate the average distortion in temporal dimension at the eye fixation level, based upon which the smoothed distortion value is recursively calculated and consolidated by the characteristics of temporal variations. Then, we give a detailed solution of how to to integrate the three existing spatial VQA metrics into our approach. Besides, the cross-format omnidirectional video distortion measurement is also investigated. Finally, the spatiotemporal distortion of the whole video sequence is obtained by pooling. Based on the modeling approach, a full reference objective quality assessment metric for omnidirectional video is derived, namely OV-PSNR. The experimental results show that our proposed OV-PSNR greatly improves the prediction performance of the existing VQA metrics for omnidirectional video.**

***Index Terms*—Omnidirectional video, objective video quality assessment, spatio-temporal distortion.**

I. INTRODUCTION

Omnidirectional video, also dubbed 360-degree video, is an emerging multimedia representation, which can provide a whole spherical space field of view (FoV) [1]. When watching a 360-degree video, the viewer usually needs to wear a head-mounted display (HMD) to freely look around through the movement of his/her head, to gain immersive and interactive experiences. Unlike two-dimensional (2D) planar video, omnidirectional video recording process can be divided into three major steps: capturing, stitching, and projection. Typically, omnidirectional video is recorded using multiple cameras or a dedicated camera that contains multiple camera lenses, to capture scene information in all directions simultaneously. Each camera lens corresponds to a separate video file, and then these footages are merged into one spherical video piece through the stitching method. The resulting video generally needs to be mapped to a plane format called panoramic video for the convenience of encoding and transmission. In addition, in order to provide realistic immersive video content, omnidirectional video usually requires ultra high definition (UHD) resolution or even higher. Due to the distortion potentially introduced by plane mapping and the extraordinarily high resolution visual content, traditional 2D video quality-related solutions are not well suitable for omnidirectional video. Therefore, there is an urgent need of a new approach for omnidirectional video quality assessment.

Video quality assessment (VQA) is of fundamental importance for optimization of the associated algorithms designed in a variety of video processing fields, such as acquisition, compression, enhancement, restoration, and transmission. Basically, VQA methods can be categorized into two classes: subjective methods and objective methods. Subjective VQA metrics measure the quality of the video by asking a number of human observers to rate scores and a mean opinion score (MOS) is computed for each video. Subjective VQA metrics may be considered the most accurate and reliable way for assessing the video quality, since it directly expresses the feeling of the viewer about the quality of the visual content. But it is time and resource consuming. Further, the mood of the subjects and the environment may affect the consistency of the results. On the other hand, the objective VQA is to design a mathematical model that can automatically approximate the evaluation results from subjective VQA. According to the availability of reference video, the objective VQA metrics can be classified as full-reference (FR), reduced-reference (RR), and no-reference (NR) metrics [2]. In this paper, we focus primarily on the full-reference objective VQA for omnidirectional video.

For conventional 2D video, Peak Signal-to-Noise Ratio (PSNR) is the most widely used FR objective VQA metric, because of its simplicity and interpretability. Another well-known 2D FR objective VQA metric is Structural Similarity Index (SSIM) [3] and its variants [4]–[6]. However, omnidirectional video is usually produced and stored in planar representations to be compatible with the existing planar video coding standards. Two sphere-to-plane projections are widely used for omnidirectional video nowadays, namely, equirectangular projection (ERP) and cube map projection (CMP). ERP projects the sphere to a plane forming a panoramic image. In this projection, a constant sampling density is used vertically on the sphere, while, horizontally, each latitude is stretched to a unit length to fit in a rectangle. CMP places the sphere at the center of a cube with unit length sides. Each face of the cube is generated by a rectilinear projection with a 90° field of view in horizontal and vertical directions. While, at the time of rendering and display, the planar representations are

P. Gao and P. Zhang are with College of Computer Science and Technology, Nanjing University of Aeronautics and Astronautics, Nanjing, China.

A. Smolic is with School of Computer Science and Statistics, Trinity College Dublin, Ireland.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3044458, IEEE Transactions on Multimedia

then mapped back to the sphere space. Due to the different formats between representation space and observation space in omnidirectional video, objective VQA becomes extremely difficult compared to 2D planar video. If we directly use conventional 2D VQA approaches as the quality measures for the planar representation of omnidirectional video, the scores obtained may not correspond well with human perception in the spherical space. Under this circumstance, several works [7]–[12] for VQA on omnidirectional video have emerged. Nevertheless, among these approaches, only the impact of spatial spherical characteristics on the quality of omnidirectional video has been studied, and the temporal distortion is ignored, which is however also crucial for VQA.

A simple and straightforward approach to implement VQA metrics is to apply an image quality assessment (IQA) metric on a frame-by-frame basis. That is, the quality of each frame is evaluated independently, and then the video global quality is obtained by using a simple average or Minkowski summation. However, unlike still images, videos are perceived by human eye in spatial as well as temporal dimensions. Thus, the representation of distortions in video should contain both spatial and temporal aspects. As a consequence, a more sophisticated modeling approach needs to be designed to take into account both spatial and temporal distortions. For 2D video VQA, a large number of works considering the temporal distortions have been proposed. However, all currently existing VQA methods for omnidirectional video only take into account spatial degradation effects, and none of them consider the effect of temporal distortions.

In this paper, we propose a new modeling approach to objectively evaluate the quality of omnidirectional video by considering both the spatial characteristics and the temporal variation of distortions across frames. Accordingly, we propose a full reference objective video quality assessment method, namely Omnidirectional Video PSNR (OV-PSNR). To be more specific, firstly, to simulate temporal change of the areas in the visual field, we construct a spatial-temporal tube based-structure on the ERP format, within which the spatial spherical distortion is measured and temporally filtered. Secondly, we calculate the gradient of the temporal distortion to evaluate the most perceptually important temporal variation of distortion at the eye fixation level. Finally, we derive the spatio-temporal distortion for the whole video sequence.

In summary, the key contributions of this paper are as follows.

- We present a spatio-temporal distortion modeling approach to evaluate the quality of omnidirectional video. Through constructing the spatio-temporal tube as a basic quality assessment unit and characterizing temporal variations of distortions across frames, a more accurate VQA modeling approach is devised for omnidirectional video.
- We develop a new FR objective VQA metric for omnidirectional video, for short, OV-PSNR, which integrates the current most popular three existing objective VQA methods for omnidirectional image/video into our spatio-temporal distortion modeling approach. We also adapt the proposed OV-PSNR for the quality evaluation to the case that the test video has a different projection format than the reference video. We demonstrate that the proposed objective VQA metric can significantly improve the performance of existing VQA metrics for omnidirectional video.

In order to make reproducible research, the implementation code of the proposed OV-PSNR in this paper is made publicly available in this repository[1].

This paper is organized as follows. In section II, we will introduce the related work. Section III describes our proposed VQA metric and the underlying spatio-temporal distortion model. Section IV presents the experimental results and related analysis. Finally, conclusion is given in Section V.

II. Related Work

A. *Objective VQA for omnidirectional video*

For traditional 2D video, a common practice to implement objective VQA metrics is to apply IQA methods to each frame independently and calculate the average value over the scores of all frames in the whole video sequence. The most commonly used IQA metric is PSNR. PSNR is calculated based on the mean squared error (MSE) between the reference and impaired signals, and it has clear physical meanings and computational simplicity. However, MSE and PSNR are criticized for not correlating well with subjective visual quality perceived by the human visual system (HVS), especially when the noise is not additive. Structural similarity (SSIM) [3] developed by Wang *et al.* is another popular method for quality assessment of still images which estimates perceptual distortions by considering structural information, and its extension MultiScale-SSIM (MS-SSIM) [4] provides more flexibility than the single-scale version by incorporating the variations of image resolution and viewing conditions. However, these methods do not utilize temporal characteristics in their HVS model benefiting for VQA. In [5], the author preliminarily extended the SSIM metric to the video domain. The Speed SSIM [6] is also proposed, which uses the SSIM index in conjunction with statistical models of visual speed perception. In addition to the PSNR and SSIM based metrics, a VQA algorithm called video quality metric (VQM) [13] from the National Telecommunications and Information Administration (NTIA) was adopted by the American National Standards Institute (ANSI) as a national standard and as a recommendation of the International Telecommunication Union (ITU), due to its excellent performance in the Video Quality Experts Group (VQEG) Phase II validation tests. In [14], the authors proposed the variation of spatial quality in time as the measure of quality fluctuation in VQA.

Unlike conventional 2D VQA, in order to evaluate the quality of omnidirectional video reasonably, it is necessary to consider the spatial stretching effects caused by the projection between spherical and planar spaces. The ERP format oversamples the sphere at the poles, resulting in stretched top and bottom areas on the ERP picture. When using CMP format, spherical positions corresponding to the center of a CMP face are sampled more sparsely compared to those corresponding to

[1] https://github.com/I2-Multimedia-Lab/360-video-experimental-platform

the sides of the face. In the latest stage of standardization for 360-degree video coding, several metrics on omnidirectional video quality assessment taking into account the mapping between representation space and observation space have been introduced. As one of the earliest works, Yu *et al.* [7] proposed a sphere-based peak signal-to-noise ratio (S-PSNR), which calculates PSNR based on uniformly sampled point set on spherical surface instead of 2D plane, and the sample values are calculated by the corresponding neighboring samples in original projection plane via the nearest neighbour or bicubic interpolation methods, generating two variants i.e., S-PSNR-NN and S-PSNR-I [15]. Additionally, Zakharchenkoa *et al.* [8] proposed a craster parabolic projection PSNR (CPP-PSNR) to convert another projection format to craster parabolic projection (CPP) plane, and calculate PSNR based on resampled points in CPP domain, where the resampled values are obtained by interpolation as well. Both VQA methods for omnidirectional video apply the interpolation algorithm to obtain the sample values, which may bring inaccuracy to evaluation results. Sun *et al.* [9] proposed the weighted-to-spherically-uniform PSNR (WS-PSNR) metric, which uses all samples on the original projection plane and considers the weights according to the corresponding specific point positions on the spherical surface, but WS-PSNR cannot assess quality of omnidirectional video across different projection formats. Similarly, an area weighted spherical PSNR (AW-SPSNR) [10] is proposed, which can utilize all available samples in the projected 2D plane and does not rely on interpolation. In [16], a Voronoi-based objective quality model is proposed for omnidirectional video quality evaluation, in which the interactive look around nature and the spherical representation characteristic are taken into account. In [17], the Video Multimethod Assessment Fusion (VMAF) is used to measure the quality of 360VR sequences. In [18], visual attention is incorporated into VQA of 360° video. In this approach, each 360° image is firstly subdivided into multiple planar patches, and then the objective quality of each patch is analysed based on visual attention. Since the main difference between 2D and omnidirectional videos is that observers only can access the content inside the FoV in omnidirectional video, there are a number of objective VQA methods based on viewers region of interest (ROI) proposed recently. In [11], Yang *et al.* proposed a VQA method based on multi-level quality factors, which calculates the panoramic video quality with region of interest (ROI) maps. Xu and Li [12] proposed the content-based perceptual PSNR (CP-PSNR), considering the possible viewing directions trained and predicted on the video contents. Recently, there have emerged several deep learning based approaches for quality evaluation on 360° contents. In [19], Lim *et al.* proposed a 360° image quality assessment method using adversarial learning. Li *et al.* [20] proposed a viewport-based convolutional neural network approach for VQA on 360°. This network firstly uses a viewport proposal network to propose potential viewport, and then employs a viewport quality network for quality prediction. However, all these mentioned metrics only take into account spatial degradation effects, and, nevertheless, temporal effects are also essential to perform quality evaluation for omnidirectional video.

B. Temporal effects in objective VQA

For objective VQA, a more sophisticated modeling approach needs to be designed since the frame-level averaging of spatial quality alone is not sufficient. At the early stages of 2D VQA research, many studies [21]–[27] have been proposed using models of the HVS to evaluate the spatio-temporal distortion. In [22], van den Branden Lambrecht has presented a complete spatio-temporal model of the HVS by considering three aspects of vision: *transient* and *sustained* temporal mechanisms, *contrast sensitivity*, and *visual masking*. Particularly, it is believed that two temporal filters need to be implemented to model the sustained and transient mechanisms of temporal vision, one low-pass and one band-pass. For example, the Moving Pictures Quality Metric (MPQM) [23], the Perceptual Distortion Metric (PDM) [26], and the Digital Video Quality (DVQ) metric [27], these works utilize two filters or a single low-pass filter to model the temporal mechanisms. Moreover, motion information also plays an important role in the perception of video quality. In [25], the Motion Rendition Quality Metric (MRQM) was proposed, which is based on an extension of the previous spatio-temporal model by incorporating human motion sensing. SSIM was extended to the temporal dimension by using a weighting scheme that takes into account motion information in [5] and [6]. In [28], the TetraVQM method estimates the motion trajectories on the reference sequence by applying block motion estimation and utilizes the motion vectors to generate the spatio-temporal distortion map. Further, Seshadrinathan *et al.* proposed [29] a general framework for measuring both spatial and temporal video distortions and a VQA algorithm called the MOtion-based Video Integrity Evaluation (MOVIE) index, which is based on their earlier work [30]. In addition to the models of visual motion sensors, there are several works that consider the temporal variations of the spatial distortions and develop a more sophisticated pooling strategy. In [31], the temporal pooling mechanism used to model continuous perceived quality recordings is introduced. In [32], an objective VQA method based on a short-term and a long-term temporal pooling is proposed. In the short-term stage, the video sequence is divided into spatiotemporal segments, which can evaluate the quality of the temporal distortions at eye fixation level and per-frame quality scores are obtained. Then the long-term stage computes the quality score for the whole video sequence by combining the average of all frame distortion scores with the temporal variation of distortions over the whole sequence.

However, to our best knowledge, there are few objective VQA metircs that take temporal mechanism into account for omnidirectional video. Therefore, we propose a spatio-temporal distortion modeling approach that the existing VQA metrics for omnidirectional video can be easily integrated with. Our analysis shows that the performance of VQA metrics for omnidirectional video can be improved significantly by the introduction of our temporal distortion model. A preliminary study of this work has been presented in [33], in which an objective model is proposed for VQA on 360° video by considering temporal aspect of perceived distortion. However, this model is only applicable to the spatial distortion in a panorama frame that is calculated using WS-PSNR. In particular, this model is only limited to the case that the reference and impaired 360° videos have the same projection format, i.e., ERP format. In this paper, we address the problem of quality evaluation for omnidirectional video thoroughly. Firstly, we propose a more sophisticated and comprehensive spatial-temporal approach for modeling the temporal distortion variations in omnidirectional video, in which a module called spatial distortion map generation is designed for adapting to various spherical distortion calculation methods. Secondly, we extend the most popular three spatial-spherical-quality-focused omnidirectional video metrics from spatial domain to temporal domain. The used three metrics are also the ones that JVET recommended for 360 video quality evaluation. Finally, cross-format omnidirectional video quality assessment is considered. Besides these improvements, we also provide a comprehensive experiment to validate the proposed spatial-temporal model.

III. The Proposed Video Quality Assessment Method

In this section, we develop an objective VQA method for omnidirectional video. Firstly, we give a brief description of how a human observer perceives a temporal distortion in Section III-A. Then, Section III-B elaborates on the proposed spatio-temporal distortion modeling approach. Finally, in order to be compatible with the three most common VQA metrics for omnidirectional video (i.e., S-PSNR, CPP-PSNR, WS-PNSR), we develop a detailed solution to incorporate them into the proposed spatio-temporal model in Section III-C.

A. Visual Attention Mechanism

In this paper, we seek to address the temporal effects in objective VQA by introducing the temporal distortions in the video sequence. A temporal distortion can be defined as the temporal evolutions of the spatial distortion, such as mosquito noise, stationary area fluctuations, jerkiness, ghosting and smearing [34]. The intense changes or fluctuations, of the spatial distortions over time can considerably influence human perception. Consequently, the question arising to know is how a human observer actually perceives a temporal distortion.

The perception of the temporal distortions is closely related to the visual attention mechanisms. In video quality evaluation, for each variation in picture quality, a stimulus is sent to the human observer, and an associated response is generated. The time frequency and the speed of the spatial distortion variations could considerably influence human perception. Generally speaking, the judgement process of quality evaluation for video conducting from human cognitive emulator can be very complicated. However, by considering the characteristics of HVS, the relationship between the temporal distortion and visual attention contains four main elements, i.e., smoothing effect, perceptual saturation, asymmetric behavior, and motion suppression [24], [31], [35]. The smoothing effect refers to that human observers integrate distortion temporally over a time window. More specifically, in a short duration, when the unimpaired frames are interleaved with the distorted ones, all frames appear distorted. Therefore, the perceived distortion at a certain frame is not just the distortion of the current frame. For perceptual saturation, it means that there are limitation in the viewer's ability to observe any further changes in the frame quality beyond certain thresholds, either toward better or worse quality. The asymmetrical behaviour is the fact that humans are better able to remember unpleasant experiences than pleasant moments, and also experience greater intensity of feelings from disliked situations compared to favourable situations. For motion suppression, since motion or temporal change is dominant features in dynamic visual scene, it increases the processing cost of visual perception and as a result of limited processing power in the HVS, greatly reduces visual sensitivity [36], [37]. Due to motion suppression, perception of distortions is considerably reduced in peripheral vision. Besides these four elements, temporal distortion may also guide strong attention to salient areas, while distortions that occur outside the salient areas are assumed to have a lower impact on the overall quality [38].

Visual perception occurs when the eyes focus on the light onto the retina, with a combination of eye movements and shifts in visual attention. The eye movements can be categorised into three major types: saccadic movements, smooth pursuit movements, and fixations [39]. Saccades are very rapid, ballistic eye movements, pursuit movements allow the eyes to closely track a moving object smoothly instead of in jumps, and fixation is the maintaining of the visual gaze on a particular area of the visual field. When a human observer assesses a video sequence, the perception of a temporal distortion is more likely to happen during a fixation or a smooth pursuit than saccades. Imagine that the video sequence can be decomposed into a series of successive spatiotemporal segments, and each segment is spatially limited by a particular area of the visual field, and temporally limited to the average duration of a fixation or a smooth pursuit. These spatiotemporal segments contain temporal variation information of spatial distortion and each segment can be used as an assessment unit to evaluate the perceived quality of spatio-temporal distortion. Further, the visual attention mechanisms indicate that the HVS integrates most of the visual information at the scale of the fixations although fixation duration is shorter than the smooth pursuit duration [39]. Therefore, the spatio-temporal distortions can be locally observed and measured during each possible fixation period. As mentioned in [32], the duration of 400 ms is chosen in accordance with the average duration of the visual fixation, which is the most simple and effective solution.

In our proposed spatio-temporal distortion model, each possible fixation can happen at a certain area of every frame in a video sequence. Therefore, we divide each frame of a video sequence into blocks and connect the related blocks along temporal domain at eye fixation level into a spatiotemporal segment, or more specifically the tube, which serves as a basic video quality assessment unit. Based on these spatio-temporal tubes, theoretical modeling of spatiotemporal distortions for omnidirectional video can be achieved. We describe it for greater details in the next subsection.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3044458, IEEE Transactions on Multimedia

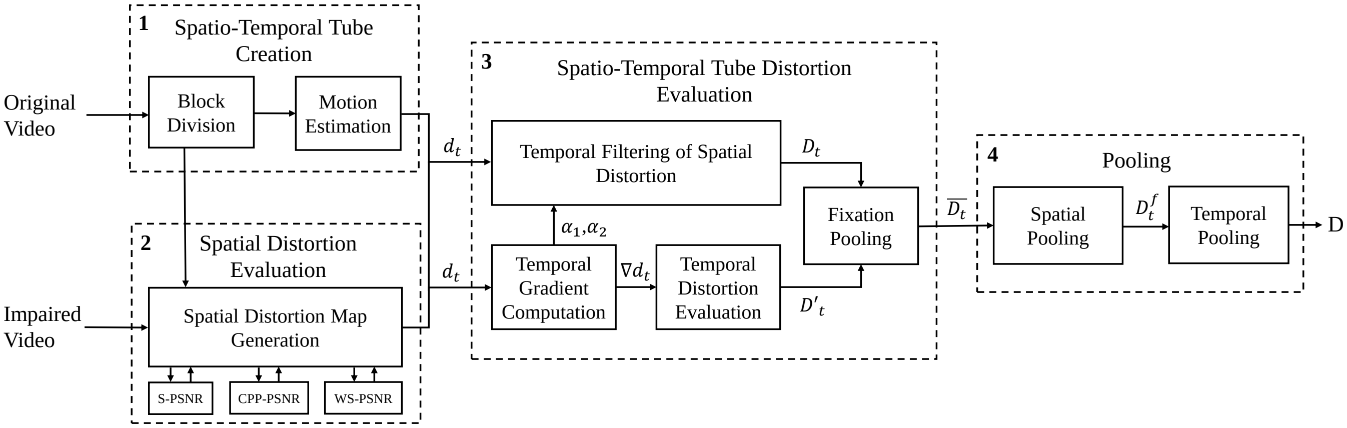

1
Spatio-Temporal Tube Creation
Original Video
Block Division
Motion Estimation
2
Spatial Distortion Evaluation
Impaired Video
Spatial Distortion Map Generation
S-PSNR
CPP-PSNR
WS-PSNR
d_t
d_t
3
Spatio-Temporal Tube Distortion Evaluation
Temporal Filtering of Spatial Distortion
D_t
α_1,α_2
Temporal Gradient Computation
∇d_t
Temporal Distortion Evaluation
D'_t
Fixation Pooling
$\overline{D_t}$
4
Pooling
Spatial Pooling
D_t^f
Temporal Pooling
D

Fig. 1: Block diagram of the proposed spatiotemporal distortion model for quality evaluation of omnidirectional video.

B. The Spatio-Temporal Distortion Model

The proposed spatiotemporal distortion model is shown in Fig. 1, which is composed of four steps. Since the spatiotemporal distortions are evaluated locally according to where visual attention positions and how long it lasts, the video sequence needs to be decomposed into a series of spatiotemporal segments. These spatiotemporal segments evaluated by a human observer during fixations can be roughly designed as spatiotemporal tubes. These tube based structures contain the spatiotemporal distortions for each possible fixation, in which a fixation can start at any frame. Therefore, the first step, *the module numbered 1* in Fig. 1, is the procedure of spatiotemporal tube creation, which is conducted in the representation space of omnidirectional video. However, in theory, it is more reasonable to construct spatiotemporal tubes in the spherical space than the considered planar space because the observation space is the endpoint where the viewer perception takes place for omnidirectional video. But block division and motion estimation on sphere is extremely challenging and still an open problem. In this paper, considering the fact that almost all the omnidirectional videos are stored and compressed in the planar format, we also focus on the ERP format to implement spatiotemporal tubes for quality evaluation. To compensate for the effect of the observation space, we will incorporate the spatial spherical characteristics into the distortion map of each block in the tube.

For a given frame F_t in the original panoramic video, we firstly divide it into $K \cdot L$ blocks, where K and L are the horizontal and the vertical number of blocks, respectively. For instance, if the block size is 16×16, an omnidirectional video with resolution of 4096×2048, there will be a total of $256 \times 128 = 32768$ blocks in each frame. Since the motion information is essential to evaluate the temporal distortion of a moving object and the locality of the temporal-corresponding block must be motion compensated, we perform the motion estimation so that the local motion between two blocks in consecutive frames is estimated. Using the estimated motion information, the past trajectory of the block can be reconstituted. More specifically, assume the number of frames in a temporal horizon is n (e.g., $n = 10$ if the fixation duration is 400 ms as mentioned above and the frame rate is 25 fps). For a block $B_{k,l,t}$ in the frame F_t, we firstly find the matching block in the previous frame for the current block using backward motion vector. The motion vectors are computed from the original sequence using block matching algorithms. There are a number of block matching algorithms in the literature, and for simplicity, we choose the New Three-Step Search Algorithm here [40]. Then, we do the similar step for the co-located block in the $t-1$ frame to find its corresponding block in the $t-2$ frame. We repeat this step until the corresponding block in the $t-n+1$ frame is found. The block $B_{k,l,t}$ and all the associated blocks in the preceding frames are connected to form a spatial-temporal tube, as shown in Fig. 2, where, meanwhile, each planar image block corresponds to an irregular area in the spherical domain. The spatiotemporal tube simulates the trajectory of a possible moving object at time and spatial domain. So far, we have built the basic quality assessment unit for omnidirectional video, i.e., the spatiotemporal tube.

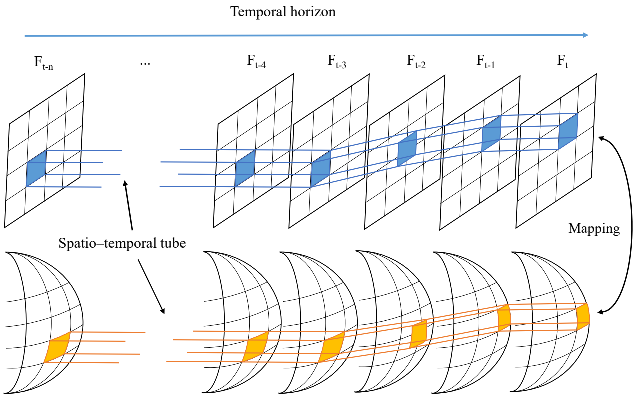


Fig. 2: Illustration of a spatio-temporal tube used to modeling the temporal distortion for omnidirectional video.

In the second step, *the module numbered 2* in Fig. 1, we do the spatial distortion evaluation for omnidirectional video to obtain the spatial distortion map for each frame F_t. The distortion $d_{k,l,t}$ of each block (k,l) obtained from the first step in this map is computed between the associated blocks in the original and the impaired frames of omnidirectional videos. Considering that the distortion map needs to reflect the spatial spherical characteristics of the omnidirectional video, this step can be achieved by extending the existing VQA metrics which take into account the effect of nonlinear projection on spatial distortion of omnidirectional video, such as S-PSNR, CPP-PSNR, and WS-PSNR, the details of which will be discussed in next subsection. After this step, it is assumed that we have a set of spatial distortion maps without temporal consideration for the omnidirectional video sequence.

The evaluation of spatiotemporal distortion for omnidirectional video is based on our spatiotemporal tube structure combined with the spatial distortion map. As shown in *the third step* in Fig. 1, the spatiotemporal tube distortion consists of two parts, the average distortion and the temporal distortion of the tube. The average distortion of the corresponding tube $D_{k,l,t}$ is not an arithmetic mean of all block distortions $d_{k,l,t}$ in the tube at the fixation duration but a smoothed average value of all block distortions considering temporal effects of HVS. To this end, a temporal filter of spatial distortion needs to be realized [31], [41]. The temporal filter acts as a low-pass filter, and uses a temporal summation in a recursive manner to convert the distortion estimate on a single frame to continuous quality estimates. Let d_t be the spatial distortion at current time t, and D_t be the smoothed (or weighted averaged) distortion at time t. It should be noted that, D_t and d_t are the simplified version of $D_{k,l,t}$ and $d_{k,l,t}$ respectively for notational conciseness. The relationship between d_t and D_t can be represented as follows:

$$D_t = \begin{cases} (1-\alpha_1)\cdot d_t + \alpha_1 \cdot D_{t-1} & \text{if } |\nabla d_t| \geq \mu \\ (1-\alpha_2)\cdot d_t + \alpha_2 \cdot D_{t-1} & \text{if } |\nabla d_t| < \mu \end{cases} \tag{1}$$

where α_1 and α_2 are the smoothing factors, ∇d_t and μ are the distortion gradient value of frame F_t and the threshold respectively, and the recursive process starts at the first block of a tube with $D_{t-n+1} = d_{t-n+1}$. If the value of $\alpha_1(\alpha_2)$ is close to zero, the less smooth effect on the current distortion D_t is from the distortion in the previous block. When $\alpha_1(\alpha_2)$ is set to 0, D_t and d_t are exactly identical. For the selection of $\alpha_1(\alpha_2)$ value, we use the distortion gradient value between adjacent frames as the indicator. If the absolute value of the distortion gradient value is greater than or equal to a threshold value μ, a larger value α_1 (e.g., 0.8) is selected, otherwise a smaller value α_2 (e.g., 0.5) will be selected. The details of distortion gradient calculation will be described in the following.

In the above, the average distortion of the spatiotemporal tube is the distortion result of temporal filtering of the spatial distortions for the block in the frame t in the corresponding tube. However, the characteristics of temporal distortions in the tube, such as frequency and amplitude of the distortion variations, also significantly impact the perception. Therefore, we consolidate the average distortion D_t with the temporal distortion D'_t which is produced by the temporal filtering of distortion variation gradient.

The distortion gradient ∇d_t at frame t is defined as follows

$$\nabla d_t = \frac{d_t - d_{t-1}}{\Delta t} \tag{2}$$

where d_t and d_{t-1} are the spatial distortions at frame t and $t-1$ in the temporal horizon of tubes, as obtained in the step 1. Δt is the time interval between frames as mentioned above. A larger value of distortion gradient means higher temporal variations which is more annoying to human observer. However, there exists a certain gradient threshold μ. If the absolute value of the distortion gradient is below μ, the temporal distortion variations hardly can be perceived. In this case, we set the associated distortion gradient to 0 to reduce the limited effect of the temporal variations.

The frequencies of the temporal variations can be represented as the number of sign changes of distortion gradients. HVS is more sensitive to temporal distortion variations at medium frequencies than at low or high frequencies [32]. Therefore, $fs(n_s)$ is a Gaussian-like fitting function as follows:

$$fs(n_s) = \frac{g_s}{\sigma_s\sqrt{2\pi}} \cdot e^{-\frac{(n_s-\mu_s)^2}{2\sigma_s^2}} \tag{3}$$

where n_s is the number of sign changes of distortion gradients, which is obtained by comparing the signs of neighbouring distortion gradients in a tube. If two neighbouring distortion gradients have different signs, we count that there is a sign change occurred. g_s represents the scaling factor, which is used to scale the probability density of standard Gaussian distribution. μ_s and σ_s represent the mean and standard deviation of the distribution, respectively. The $fs(n_s)$ function reaches its maximum in the case of only one sign change of the distortion gradients in the tube duration.

Then, the temporal distortion of a tube D'_t can be calculated as

$$D'_t = Max(\nabla d_t) \cdot fs(n_s) \cdot D_t \tag{4}$$

The function of $Max(\nabla d_t)$ is to obtain the maximum distortion gradient in the tube duration. The product of $Max(\nabla_{d_t}) \cdot fs(n_s)$ is the output of temporal filtering of distortion gradients in the tube. Further, we assume that the temporal distortion is linearly related to the average distortion D_t, and thus we multiply the filtered distortion gradient $Max(\nabla_{d_t}) \cdot fs(n_s)$ with D_t to obtain the temporal distortion as shown in (4). D'_t is the result after combining the amplitude and frequency of the distortion gradient to consolidate the average distortion D_t of the tube.

The results coming from these two branches are then mixed together in the fixation pooling step, in which the average distortion D_t and the temporal distortion D'_t are merged in order to generate the final spatiotemporal distortion of tube $\overline{D_t}$,

$$\overline{D_t} = D_t + \beta \cdot D'_t \tag{5}$$

where β is a weighting factor that gives how important the temporal variations are. In the case of $\beta = 0$, the $\overline{D_t}$ is reduced to the spatiotemporal tube distortion without consideration of temporal variations of distortions. Until now, the perceived distortion of an omnidirectional video sequence is evaluated

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.3044458, IEEE Transactions on Multimedia

at the fixation level, resulting in the final distortion for the spatiotemporal tube quality assessment unit $\overline{D_t}$, more specifically $\overline{D_{k,l,t}}$. We will pool the distortions of these units together to get a global score to evaluate the quality of the whole video sequence in the next step.

There are two steps in the pooling stage, as *numbered 4* in Fig. 1. In the first, a per-frame perceptual distortion score D_t^f from the spatiotemporal distortions of all the tubes $D_{k,l,t}$ finishing at frame t is computed. This is performed by using a Minkowski norm

$$D_t^f = (\frac{1}{K \cdot L}\sum_{k=0}^{K-1}\sum_{l=0}^{L-1}(\overline{D_{k,l,t}})^{\beta_s})^{\frac{1}{\beta_s}} \tag{6}$$

where β_s is the Minkowski exponent, and β_s = 2 works well here.

PSNR values for video sequence can be calculated in two ways. One is an average value of the per-frame PSNR, called method Av-Log. And in the second method Log-Av, PSNR is based on the average distortion of frames. The study [42] shows that PSNR of Log-Av is a better value for measuring PSNR of video sequences. Thus, in the second step of pooling, the average distortion value is calculated from the all frame distortions over the whole video sequence, called D.

$$D = \frac{1}{N}\sum_{t=0}^{N-1} D_t^f \tag{7}$$

Finally, the so-called quality assessment score OV-PSNR is derived from our spatiotemporal distortion model, which is calculated as follows

$$OV\text{-}PSNR = 10 \cdot log(\frac{MAX_I^2}{D}) \tag{8}$$

where MAX_I is the maximum possible pixel value of the color space. Generally speaking, OV-PSNR can be used to objectively measure the perceptual quality of omnidirectional video. However, there is lack of explanation for the details about how to integrate the existing VQA metrics for omnidirectional video into our spatiotemporal distortion model to generate spatial distortion maps as shown in the second step. We discuss this topic in next subsection.

C. Compatibility with the Existing VQA Metrics

A major difference of VQA metrics for omnidirectional video with 2D video is the modeling of the stretching effects caused by the projection from spherical space to planar space. In light of this, several proposed VQA metrics for omnidirectional video have addressed this aspect in different ways, and all of them achieve very reasonable spherical quality estimation performance. In this study, we attempt to integrate the three most commonly used VQA metrics for omnidirectional video, i.e., S-PSNR, CPP-PSNR, and WS-PSNR, into our distortion evaluation model as described above. Note that in the second step of our spatiotemporal distortion model, we have mentioned that a spatial distortion map needs to be generated to output each block distortion $d_{k,l,t}$ in the corresponding tube finishing at each frame. In the following, we show how the spatial distortion map is obtained by using these three different VQA methods. The detailed distortion map generating process is illustrated in Fig. 3.

1) Temporal Extension to WS-PSNR: The method WS-PSNR measures omnidirectional video quality directly in the projection domain by assigning different weights to each image samples on the 2D projection plane. Therefore, we can naturally generate the spatial distortion map for blocks by directly using the weight map produced by WS-PSNR method since our modeling approach performs block division and tube creation exactly on the 2D projection plane. Let $d_{k,l,t}$ be the spatial distortion of the block (k,l) in frame t. Thus, the $d_{k,l,t}$ can be calculated between the original video sequence and the impaired video sequence in the form of a weighted mean squared error (WMSE) [9], which considers stretching ratio of areas from the projection plane to spherical surface for omnidirectional videos, i.e.,

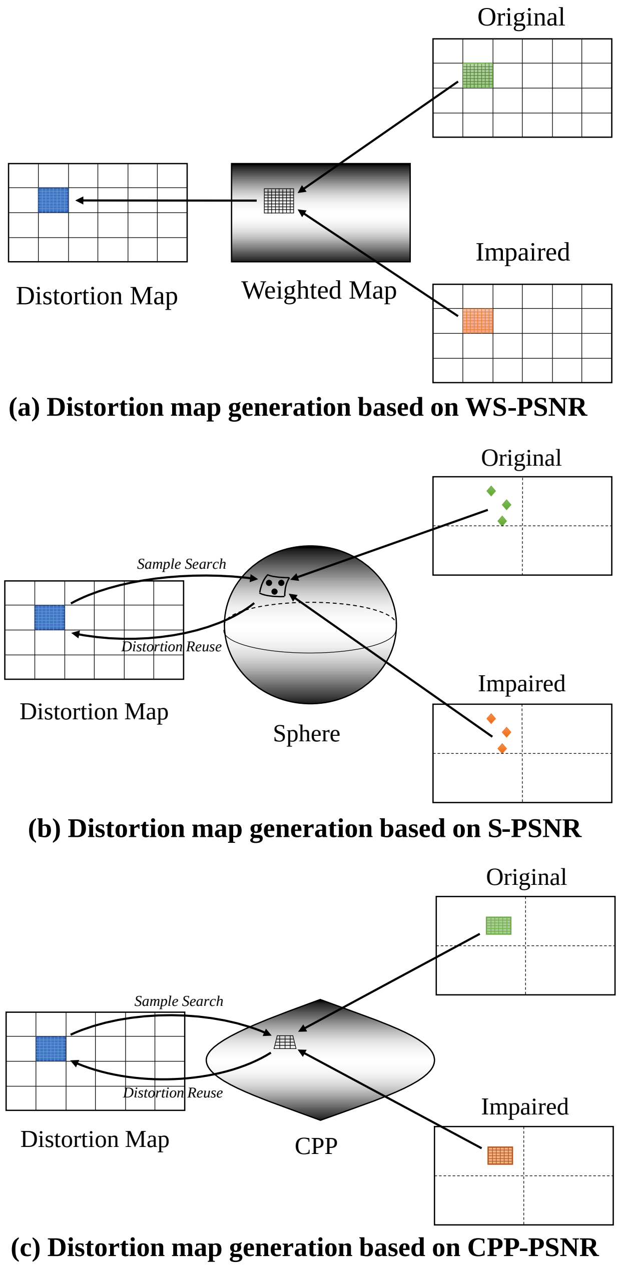
$$d_{k,l,t} = \frac{\sum\limits_{i=k\times M}^{(k+1)\times M-1}\ \sum\limits_{j=l\times N}^{(l+1)\times N-1}[(y(i,j)-y'(i,j))^2\cdot\omega(i,j)]}{\sum\limits_{i=k\times M}^{(k+1)\times M-1}\ \sum\limits_{j=l\times N}^{(l+1)\times N-1}\omega(i,j)} \tag{9}$$

where $y(i,j)$ and $y'(i,j)$ are the pixels in the block (size $M \times N$) of original and impaired video frames, respectively, and $\omega(i,j)$ is the weighting factor value which can be calculated as the stretching ratio of the area in projection format and the area in the spherical domain. When the ERP format is adopted as the projection format, the weight $\omega(i,j)$ at position (i,j) in an $W \times H$ image is calculated as

$$\omega(i,j) = cos\frac{(j+0.5-H/2)\pi}{H} \tag{10}$$

However, it should be noted here that, WS-PSNR can be only used for the case that original video and impaired video have the same resolution and the same projection format. Thus our modeling approach combining with the WS-PSNR method also has this limitation. Since the original video that is used for building up our spatio temporal model requires to be the ERP format as mentioned earlier, when extending the WS-PSNR metric to our temporal model, the projection format of the test video (i.e., the corresponding impaired video) is also limited to the ERP format.

2) Temporal Extension to S-PSNR: The computation procedure of spatial distortion by extending S-PSNR is generally more complicated than by extending WS-PSNR. Since such VQA metric can support omnidirectional video quality distortion measurement across projection formats, an intermediate layer (i.e., spherical domain for S-PSNR) usually exists for distortion computation from the ground truth signal. Therefore, we can generate the spatial distortion map for planar blocks by two steps, i.e., searching the sample in the spherical domain and reusing the distortion result of the sample in the intermediate domain. In a typical S-PSNR process, it is known that a set of points has been uniformly pre-sampled on a sphere for the purpose of discretizing the signal on the spherical space. Therefore, in the first step, we map the pixels of each block of the ERP plane to the spherical surface and determine the pre-sampled points contained in the corresponding irregular



(a) Distortion map generation based on WS-PSNR

(b) Distortion map generation based on S-PSNR

(c) Distortion map generation based on CPP-PSNR

Fig. 3: Generating the spatial distortion map for blocks by using WS-PSNR, S-PSNR, CPP-PSNR.

area. Then, the average distortion of the matched pre-sampled points can be regarded as the distortion value $d_{k,l,t}$ of the relating block on the 2-D plane, as shown in Fig. 3(b).

$$d_{k,l,t} = \frac{\sum_{i=0}^{N-1} [(y(i) - y'(i))^2]}{N} \tag{11}$$

where $y(i)$ is the signal value of pre-sampled point i on spherical domain and $y'(i)$ is its reconstructed signal value. N is the number of pre-sampled points contained in the mapped area. It should be noted here that, the values of $y(i)$ and $y'(i)$ can be obtained by further projecting the pre-sampled points on the sphere to the original and impaired frames. In this projection, the mapped position on the original and impaired frame may not be in the integer position. In this case, as illustrated in S-PSNR, the nearest neighbouring interpolation may be used. Interested readers are referred to the S-PSNR-NN [15] for more details.

Having outlined the basic idea of obtaining the spatial distortion map by re-using the distortion of S-PSNR, we present the detailed procedure of establishing the relationship between our planar blocks and S-PSNR pre-sampled points on sphere. As mentioned before, this paper adopts the ERP format as reference video format so that the block distortion is always calculated on a planar map, but the spherical points in S-PSNR are sampled with longitude and latitude coordinate system. Commonly, a longitude λ is in the range $[-\pi, \pi]$ and a latitude ϕ is in the range $[-\pi/2, \pi/2]$. To realize the sample search step as mentioned above, a planar block needs to be converted to spherical coordinates for the determination of which pre-sampled point is the best matching point for this pixel. The coordinate conversion from a 2D position (i, j) to (λ, ϕ) in the ERP can be achieved by using

$$\lambda = (\frac{i}{W} - 0.5) \times 2\pi \tag{12}$$

$$\phi = (0.5 - \frac{j}{H}) \times \pi \tag{13}$$

However, it is impossible to have exact one-to-one match between points mapped from a block and spherical pre-sampled points. To this end, we use the nearest pre-sampled points to the mapped points as the final sample search result. Further, for the calculation convenience, we introduce XYZ coordinate system, where the (X, Y, Z) coordinates on the unit sphere can be transformed from (λ, ϕ) using [43]

$$X = \cos(\phi)\cos(\lambda) \tag{14}$$

$$Y = \sin(\phi) \tag{15}$$

$$Z = -\cos(\phi)\sin(\lambda) \tag{16}$$

Then, the spherical pre-sampled points required for calculating block distortion in (11) are determined by comparing the Euclidean distance between the XYZ coordinates of pre-sampled points and those of the mapped points. It is worth noting that, directly comparing Euclidean distance for a mapped point with a pre-sampled point on a sphere (around a total of 655362 points) would induce intractable complexity. Here, we use k-dimensional tree (or K-D tree) [44] to improve the computational efficiency. A K-D tree is a binary search tree for organizing points in a k-dimensional space which is a useful data structure involving a multidimensional search key. After the relationship between the planar blocks and the spherical pre-sampled points is determined, the distortion of the pre-sampled points can be derived from the S-PSNR-NN method process. Finally, we use the S-PSNR-NN calculated distortion as the spatial distortion for each pixel and do arithmetic mean to yield the distortion of each corresponding planar block.

3) Temporal Extension to CPP-PSNR: Similar to S-PSNR, our spatiotemporal model with CPP-PSNR also uses an intermediate domain (i.e., CPP domain) to obtain the distortion map of blocks. CPP plane considers equal point distribution on a sphere and preserves constant spatial resolution. To generate the spatial distortion map needed in our model, we do the

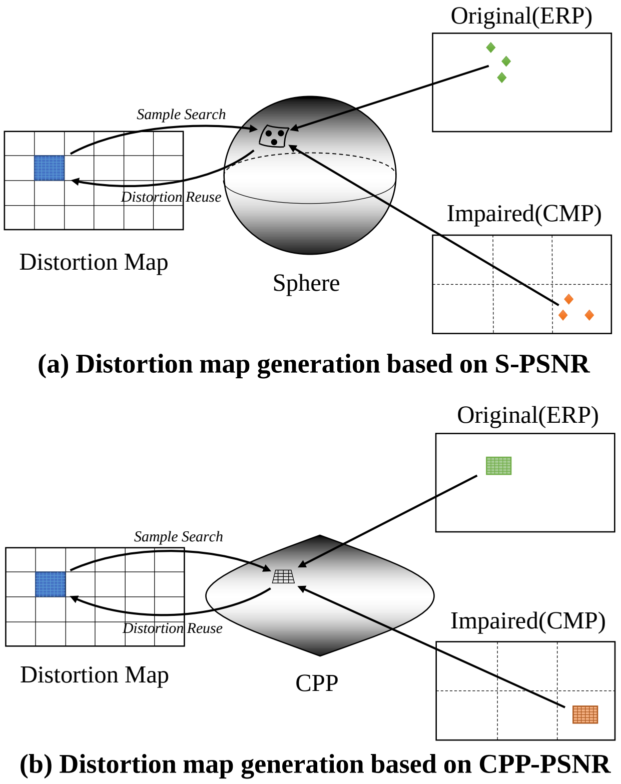


Fig. 4: Our modeling approach with (a) S-PSNR or (b) CPP-PSNR supports the distortion measurement between different projection formats. Note that the original input here must be the ERP format, and the impaired input can be another projection format, e.g., CMP 3×2 type.

sample searching and distortion result reusing in the CPP domain, as shown in Fig. 3(c). Firstly, we transform position (i, j) of points in each block on 2D ERP plane to (λ, ϕ) coordinate system by using the equations of (12) and (13), and then map these coordinates to CPP plane by using the following

$$m = W_{cpp} \times \{\frac{R\lambda}{2\pi}[(2\cos\frac{2\phi}{3}) - 1] + 0.5\} \tag{17}$$

$$n = H_{cpp} \times [0.5 - R\sin\frac{\phi}{3}] \tag{18}$$

where (m, n) is the corresponding position on the CPP plane with respect to the point (i, j) in the block of the tube, and R represents the radius of the sphere, which is equal to 1 for unit sphere, while W_{cpp} and H_{cpp} are the width and the height of CPP plane respectively. After the relationship between the planar blocks and the positions in CPP distortion map is determined, the average distortion of the corresponding points in CPP plane is calculated by using the CPP-PSNR, and then used as the result of the block distortion, similar to (11). In a typical CPP-PSNR distortion calculation, the original frame and the impaired frame should be converted into the CPP domain, and then a CPP distortion map is obtained by applying the MSE calculation.

4) Cross-format Distortion Measurement: Due to the intermediate layer used, the S-PSNR and CPP-PSNR support that the two inputs to the distortion measurement can have different projection formats. With the extension of S-PSNR and CPP-PSNR, our spatiotemporal modeling approach inherits this merit. As mentioned before, since the original source frame needs to be divided into blocks forming the tubes in the first step of our spatiotemporal model, the original video source must be the ERP format. However, the impaired source can be a different projection format, such as the CMP format. Fig. 4 shows the procedure of generating spatial distortion map between the different projection-format singal sources (ERP and CMP). Similarly, the distortion map generating processing includes the sample search and distortion reuse steps. In the distortion reuse step, the distortion of the sample on the spherical domain or CPP domain is calculated by using the S-PSNR or CPP-PSNR, respectively. As observed from Fig. 4, the areas containing the sample points corresponding to the block in the distortion map may locate at different places in the original and impaired frames. In S-PSNR, the spherical space is used to determine the sample point in the two different projection formats, while in CPP-PSNR, the CPP plane can be employed to determine the sample point in the different projection formats. Analogously, due to the existence of the intermediate domain of S-PSNR and CPP-PSNR methods, our spatial distortion map generating process can work effectively without considering whether original and impaired frames have different projection formats and interact with the intermediate domain only. In addition to the CMP format shown in the Fig. 4, Equal-area projection format (EAP), Octahedron projection format (OHP), and Icosahedron projection format (ISP) [43] also have the similar 3D-to-2D coordinate mapping processing steps to generate our spatial distortion map.

So far, we have presented the entire procedure of the proposed spatiotemporal modeling approach for omnidirectional video and the effectiveness of OV-PSNR metric will be demonstrated in the experimental section.

IV. Experimentation and Analysis

A. Dataset and Testing Procedure

In this section, we validate the performance of our proposed objective VQA method for omnidirectional video on the video dataset VR-VQA48 unless otherwise stated [12], which is publicly available online [45]. This dataset consists of 12 original omnidirectional video sequences (in YUV 4:2:0 format at the resolution of 4096×2048) and 36 corresponding impaired sequences obtained by encoding each original sequence with 3 different bitrate settings. Figure 5 shows the test sequences provided in the dataset VR-VQA48. Additionally, there are 48 subjects involved to give raw subjective quality scores for all the 48 sequences. The range of the subjective score is from 0 (lowest quality) to 100 (highest quality). Currently, two metrics are widely used in subjective VQA: one is the mean opinion score (MOS) [46] calculated by arithmetic mean over all raw subjective scores for each sequence; and the other is the difference MOS (DMOS) [47], which is the difference between a MOS for reference video and a MOS for impaired video. In this work, we calculate the DMOS as the subjective ground truth. Specifically, the DMOS is calculated as follows.

Firstly, let S_{ij} and S_{ij}^{ref} denote the raw subjective scores assigned by subject i to video sequence j and its corresponding

Fig. 5: Test sequences in the VR-VQA48 dataset.

reference sequence, and the difference scores d_{ij} are calculated by

$$d_{ij} = S_{ij}^{ref} - S_{ij} \tag{19}$$

The difference scores for reference sequences are 0 and removed so that there are 36 remaining difference scores in our experiment. Afterwards, the difference scores are converted to Z-scores z_{ij}, which then are normalized and rescaled to z'_{ij}. The z'_{ij} is lied in the range [0,100]. The related equations are shown below

$$z_{ij} = \frac{d_{ij} - \mu_i}{\sigma_i} \tag{20}$$

$$z'_{ij} = \frac{100(z_{ij} + 3)}{6} \tag{21}$$

where μ_i and σ_i are the mean value and the standard deviation of the score from the subject i.

Finally, the DMOS of each sequence j is calculated as the average of the rescaled Z-scores from all the M_j subjects.

$$DMOS_j = \frac{1}{M_j} \sum_{i=1}^{M_j} z'_{ij} \tag{22}$$

To compare the performance of our VQA method with other objective VQA methods, we follow the instructions from the Video Quality Expert Group (VQEG) Phase II FR-TV Validation Test Final Report [48], measuring the correlation between objective quality scores and subjective quality scores. Three performance indicators are used as recommended by VQEG, Pearson Linear Correlation Coefficient (PLCC), Spearman Rank Order Correlation Coefficient (SROCC), Kendall Rank Order Correlation Coefficient (KROCC), Root-Mean-Square Error (RMSE), and Mean Absolute Error (MAE). The PLCC measures the prediction accuracy of an objective VQA method, while SROCC and KROCC measure the prediction monotonicity, and RMSE and MAE quantify the difference between the objective and subjective VQA results. Moreover, to remove any nonlinearity occurred in the subjective rating process and facilitate the comparison of the models in a common analysis environment, the outputs by the objective video quality methods (the Video Quality Rating, VQR) should be mapped to the subjective scores (DMOS) space by performing a nonlinear regression fitting. We apply a 3-parameters logistic regression function, as also recommended by VQEG [48], to transform the set of VQR values to a set of predicted MOS values (DMOSp), which are then compared with the actual DMOS values. The DMOSp computation is given by

$$DMOS_p = \frac{b1}{1 + e^{-b2 \cdot (VQR - b3)}} \tag{23}$$

where $b1, b2, b3$ are the fitting parameters which are obtained by the non-linear least squares optimization. Note that the DMOS value indicates the quality difference between the impaired video and the reference video, which means, the larger value, the worse quality. Therefore, we reverse DMOS values (i.e., subtracted from 100) in regression fitting for ease of comparison. Once the nonlinear transformation is applied, the prediction performance of objective VQA models are evaluated by calculating PLCC, SROCC, KROCC, RMSE, and MAE on the value sets [DMOS, DMOSp].

B. Performance of Our Objective VQA Method

In this section, we test the performance of our proposed objective metric, i.e., OV-PSNR. Depending on how the spatial distortion map is generated, our proposed OV-PSNR includes four versions, i.e., OV-PSNR[PSNR], OV-PSNR[S-PSNR], OV-PSNR[CPP-PSNR], and OV-PSNR[WS-PSNR]. Note that, OV-PSNR[PSNR] denotes the proposed temporal model combined with the distortion map that is calculated by directly using the PSNR metric for the ERP format without considering the mapping distortion. OV-PSNR[S-PSNR], OV-PSNR[CPP-PSNR], and OV-PSNR[WS-PSNR] are the proposed temporal model with extending to S-PSNR, CPP-PSNR and WS-PSNR respectively. These variants based on PSNR for omnidirectional video are recommended by JVET common test conditions and evaluation procedures for 360 video [15]. In OV-PSNR[S-PSNR] and OV-PSNR[CPP-PSNR], we use the nearest neighbour interpolation version of S-PSNR and CPP-PSNR for performance comparison test. Further, five non-PSNR-based objective VQA methods (SSIM [5], VIF [49], FSIM [50], GMSD [51], VMAF [52]) are also calculated for comparison. In our first test, we focus on the performance validation for the case that same projection format (i.e., ERP) is employed for both the original and impaired sequences. Moreover, it is necessary to mention that, the block size of the tube in our model is set to 16×16, and the values of the parameters α_1, α_2, μ mentioned in Section III-B are $0.8, 0.5, 2.5$, which is deduced empirically from experiments. The g_s, μ_s, σ_s of $fs(n_s)$ in (3) are $16, 1, 6.2$, respectively. Then we set the value β in (5) to 1.0 without bias.

Performance on VR-VQA48 dataset. Figure 6 demonstrates the scatter plots between objective and subjective results

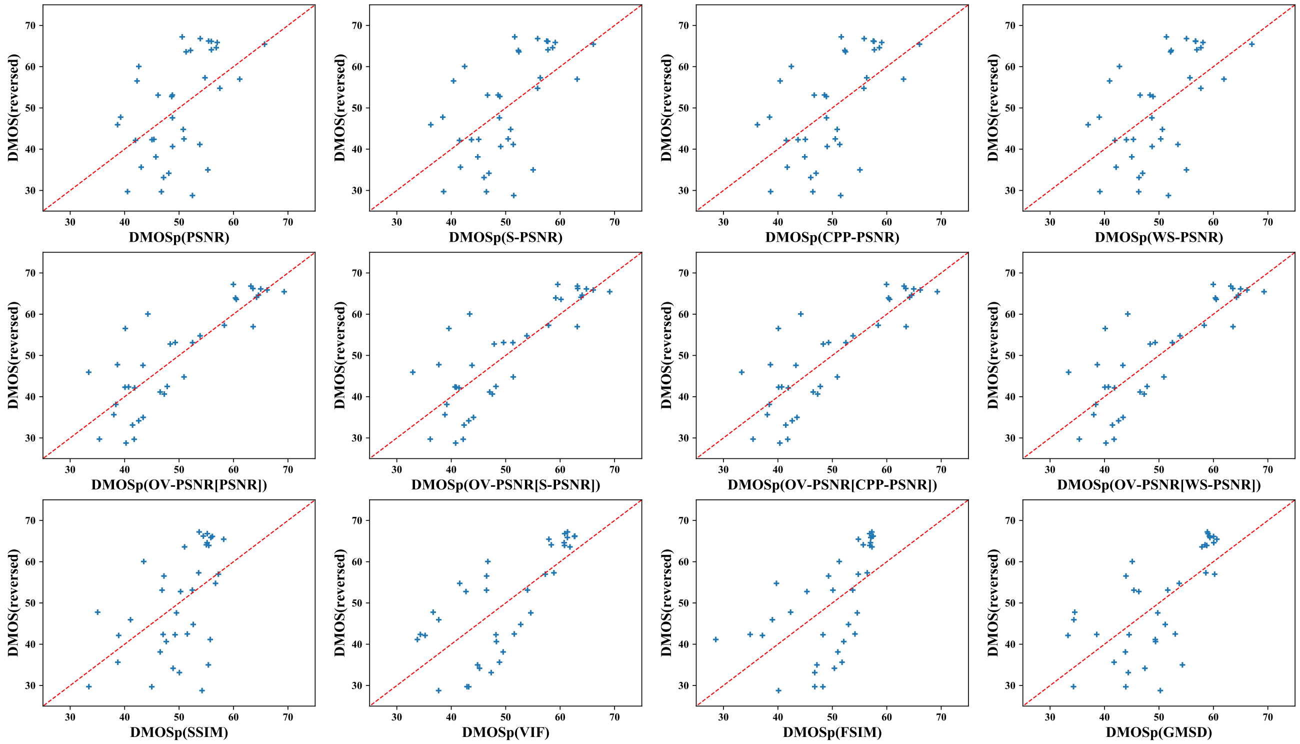


Fig. 6: Scatter plots of various objective VQA scores versus subjective DMOS for 36 impaired video sequences of VR-VQA48 dataset. Vertical and horizontal axes represent the subjective (reversed DMOS) and objective (transformed DMOSp) measurement, respectively. The second row shows the fitting performance of the proposed OV-PSNR metric, where four VQA versions of OV-PSNR are tested, i.e., OV-PSNR[PSNR], OV-PSNR[S-PSNR], OV-PSNR[CPP-PSNR], and OV-PSNR[WS-PSNR].

TABLE I
COMPARISON OF THE PERFORMANCES OF OBJECTIVE VQA METRICS ON VR-VQA48 DATASET.

Methods	PLCC	SROCC	KROCC	RMSE	MAE	TIME (s)
PSNR	0.499	0.508	0.327	10.732	9.143	0.001
S-PSNR [7]	0.569	0.595	0.384	10.183	8.539	0.347
CPP-PSNR [8]	0.567	0.595	0.381	10.198	8.551	1.986
WS-PSNR [9]	0.548	0.562	0.365	10.358	8.804	0.098
OV-PSNR[PSNR]	**0.837**	**0.790**	**0.603**	**6.749**	**5.158**	2.724
OV-PSNR[S-PSNR]	**0.818**	**0.775**	**0.584**	**7.123**	**5.505**	2.961
OV-PSNR[CPP-PSNR]	**0.837**	**0.787**	**0.600**	**6.776**	**5.181**	4.712
OV-PSNR[WS-PSNR]	**0.838**	**0.790**	**0.603**	**6.749**	**5.157**	2.902
SSIM [5]	0.506	0.532	0.378	10.679	9.051	2.256*
VIF [49]	0.722	0.721	0.562	8.565	7.667	4.802*
FSIM [50]	0.573	0.728	0.559	10.151	9.198	3.122*
GMSD [51]	0.672	0.708	0.495	9.170	7.831	0.751*
VMAF [52]	0.783	0.771	0.568	7.712	6.601	0.507
NCP-PSNR [12]	0.725	0.702	N/A	8.539	6.770	0.025*
CP-PSNR [12]	0.764	0.751	N/A	7.991	6.657	2.405*

* methods are implemented in MATLAB, while others are implemented in C++.

for 36 impaired sequences of VR-VQA48 dataset. It can be noticed that the data points from our method (in the second row) are less scattered than those of other methods, and scatter points more close to the straight line $y = x$ means a higher linear relationship with subjective quality judgments. As shown in Table I, the quantification results are reported about the PLCC, SROCC, KROCC, RMSE, and MAE between the DMOSp (from four existing PSNR-based omnidirectional video-specific metrics, five traditional 2D video quality metrics, and our corresponding OV-PSNR version metrics) and reversed DMOS. It can be clearly seen that three versions (OV-PSNR[PSNR], OV-PSNR[S-PSNR], OV-PSNR[WS-PSNR]) of our processed method perform much better than the corresponding PSNR methods without considering temporal distortion. Among them, the OV-PSNR[WS-PSNR] version achieves the best performance. This may be because, WS-PSNR measures the quality from the original signal so that our spatiotemporal quality assessment units (tubes built on ERP plane) can use all the true samples on ERP plane. On the contrary, other methods generally involve intermediate layer and interpolation filter, which may degrade the estimation accuracy to some extent. Additionally, the results of OV-PSNR[PSNR] and OV-PSNR[WS-PSNR] are so close, and we believe, this is due to the fact that the temporal factor has a greater impact on performance than spatial factors for omnidirectional video quality assessment.

At the bottom of Table I, we also include the results of NCP-PSNR and CP-PSNR measured on the same dataset VR-VQA48 for better performance comparison. It should be noted that, since the source codes of these two metrics are not available, the data of NCP-PSNR and CP-PSNR are directly taken from their original paper, i.e., Xu *et al.*'s paper [12], and the experimental environment that produces these data is slightly different from ours. As can be seen, all versions of our proposed OV-PSNR metric perform better than these two latest omnidirectional video quality metrics.

In Table I, we also compare the time complexity of the proposed metric with other methods. The experiment is performed on a desktop equipped with Intel® Core^{TM} I7-6700 and 8G RAM memory. All the test methods run on an omnidirectional video with the resolution of 4096×2048. The time shown in the table represents the running time per frame in seconds for each method. As can be observed, each temporal extension metric OV-PSNR consumes more time than its corresponding original spatial distortion modeled metric. The major time increment mainly comes from the modules of spatio-temporal tube creation, the spatial distortion map generation, and the calculation of spatio-temporal tube distortion. As each block in each frame needs to create a tube and this process involves into *motion estimation*, the tube creation part is the most time consuming one in our modeling approach. In the test, we found that it takes a portion of 50%∼90% of the overall execution time. In the time-constrained scenario, to facilitate the use of our proposed approach, we can choose to create the spatio-temporal tubes offline. Although our proposed OV-PSNR metric increases many complexity extending from the spatial distortion model, it still achieves about the same amount of complexity as other metrics, e.g., VIF, CP-PSNR, etc. This demonstrates the superiority of our proposed spatio-temporal approach, i.e., it yields the best quality evaluation performance with modest computational complexity increment.

TABLE II
COMPARISON OF THE PERFORMANCES OF OBJECTIVE VQA METRICS ON VQA-ODV DATASET.

Methods	PLCC	SROCC	KROCC	RMSE	MAE
PSNR	0.629	0.630	0.469	8.141	6.339
S-PSNR [7]	0.655	0.657	0.457	7.919	6.107
CPP-PSNR [8]	0.658	0.660	0.457	7.885	6.085
WS-PSNR [9]	0.640	0.638	0.437	8.049	6.235
OV-PSNR[PSNR]	0.735	0.730	0.520	7.100	5.577
OV-PSNR[S-PSNR]	0.727	0.720	0.510	7.190	5.599
OV-PSNR[CPP-PSNR]	0.734	0.729	0.517	7.110	5.583
OV-PSNR[WS-PSNR]	0.735	0.731	0.520	7.100	5.578
BP-QAVR [11]	0.659	0.680	0.478	8.911	7.082
VR-IQA-NET [19]	0.371	0.338	0.226	10.998	9.101
V-CNN [20]	0.874	0.896	0.713	5.755	4.489

Performance on VQA-ODV dataset. In order to further validate the performance of the proposed modeling approach, we conduct the objective experiment on another 360 degree video dataset, i.e., the VQA-ODV dataset developed in [53]. This dataset contains 60 reference sequences, 540 distorted sequences (432 impaired omnidirectional video sequences for training and 108 impaired sequences for test), and associated DMOS [54], which is known to be the largest VQA dataset for 360 degree video currently. The scatter plots of the objective VQA results versus the DMOS values on this dataset are illustrated in Fig. 7, where we apply the logistic function used in [20] for objective score fitting. In the experiment, the parameter settings of our OV-PSNR models are the same with previous experiment on VR-VQA48 dataset. As can be observed from the figure, all the proposed OV-PSNR metrics can generally better fit the ground truth DMOS compared to other methods, which demonstrates that the proposed OV-PSNR metrics have a higher correlation with the subjective DMOS results. The performance validation results of the proposed model and other metrics on this dataset are tabulated in Table II, where we also compare the proposed OV-PSNR model with several deep learning based omnidirectional video VQA methods, i.e., BP-QAVR [11], VR-IQA-NET [19], and V-CNN [20]. In this comparison, the result for BP-QAVR is obtained by re-training the model on the VQA-ODV dataset, while the results for the other two methods are obtained by directly evaluating the pre-trained models provided by the authors on the test sequences. As can be observed from the table, our proposed modeling approach can generalize well on this large-scale dataset, in which, for example, the OV-PSNR variant OV-PSNR[WS-PSNR] achieves the PLCC of 0.735 and SROCC of 0.731. Besides, our proposed approach significantly outperforms BP-QAVR and VR-IQA-NET. However, compared to the latest V-CNN, our proposed approach exhibits worse VQA performance. It should be noted that, the V-CNN is based on *deep learning* with two very complicated convolutional neural networks, one for viewport proposal and another for VQA score rating, which has also already been trained on this dataset. If one uses the V-CNN model to evaluate the quality of other omnidirectional video dataset, it needs a considerable amount of time for re-training. In addition, there is a large number of hyper-parameters to be tuned during training/re-training. The evaluation performance would be easily changed if one fails to choose one appropriate hyper-parameter. In contrast, our proposed quality assessment approach that purely relies on *mathematical modeling* can be easily deployed in practice for any dataset. Moreover, our proposed method can be compatible well with the existing spatial distortion based quality metrics of 360 degree video, e.g., WS-PSNR, S-PSNR, and CPP-PSNR.

TABLE III
CROSS DATASET VALIDATION OF OUR METHODS.

VQA-ODV → VR-VQA48					
Methods	PLCC	SROCC	KROCC	RMSE	MAE
OV-PSNR[PSNR]	0.851	0.790	0.603	6.959	5.733
OV-PSNR[S-PSNR]	0.840	0.775	0.584	7.176	5.910
OV-PSNR[CPP-PSNR]	0.850	0.787	0.600	6.977	5.746
OV-PSNR[WS-PSNR]	0.851	0.790	0.603	6.959	5.733
VR-VQA48 → VQA-ODV					
Methods	PLCC	SROCC	KROCC	RMSE	MAE
OV-PSNR[PSNR]	0.724	0.731	0.520	7.514	5.973
OV-PSNR[S-PSNR]	0.712	0.720	0.510	7.627	5.982
OV-PSNR[CPP-PSNR]	0.723	0.729	0.517	7.522	5.974
OV-PSNR[WS-PSNR]	0.724	0.731	0.520	7.514	5.973

Cross dataset validation. To validate the generalization ability and robustness of our proposed methods and the logistic fitting procedure, we conduct a cross dataset experiment. Specifically, we use the DMOSp function regressed from the VQA-ODV dataset to predict the MOS values of the proposed approach on VR-VQA48 dataset, or vice versa. For simplicity, these two test scenarios are represented as VQA-ODV→VR-VQA48 and VR-VQA48→VQA-ODV, respectively. Table III lists the performance results of cross dataset validation.

As shown in Table III, all our designed OV-PSNR models achieve very good performance, regardless of the dataset and the logistic fitting parameters, verifying the robustness and generalization ability of the proposed OV-PSNR model.

C. Influence of Different Parameter Settings

To validate the necessity of every part and compare the performances of different parameters in our spatiotemporal

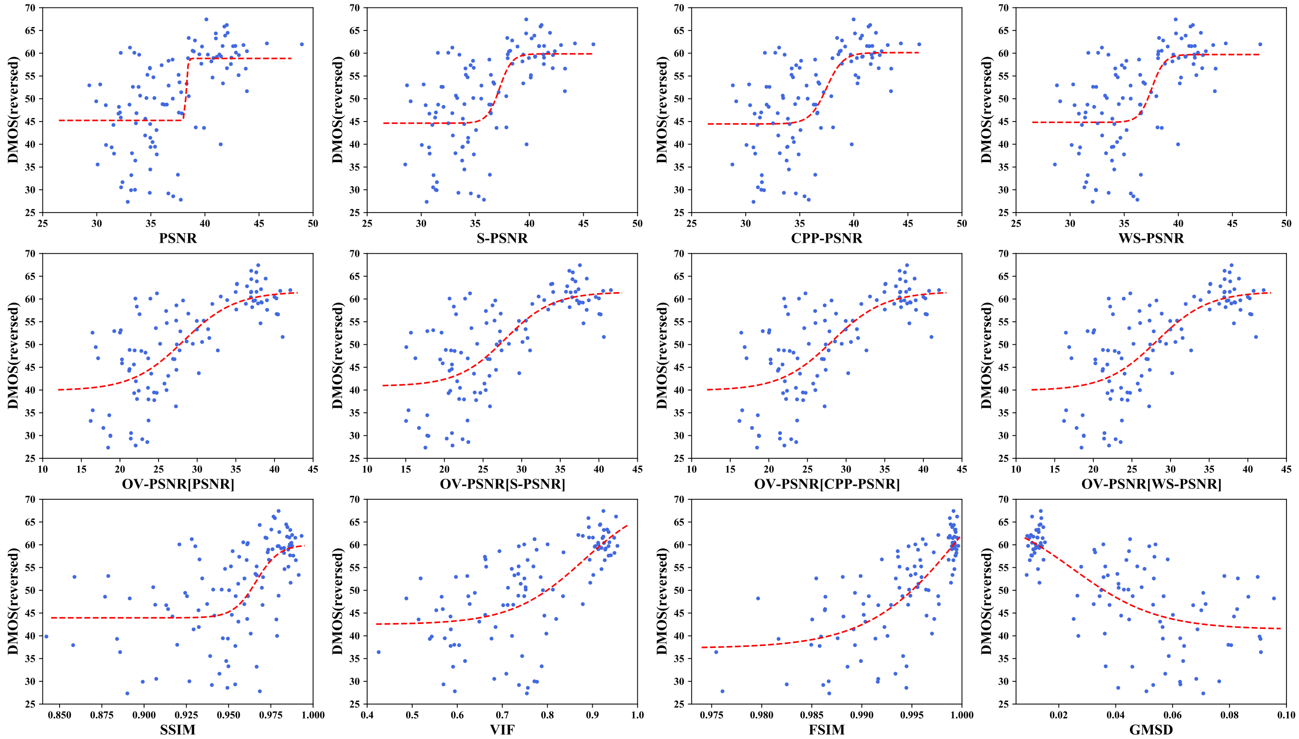


Fig. 7: Scatter plots for the objective VQA scores versus the related reversed DMOS over 108 impaired sequences of VQA-ODV dataset. Vertical and horizontal axes represent reversed DMOS and objective VQR scores, respectively. The logistic fitting curves are also shown in the figure.

model, we design a test by setting different values to the free variables of our model. As described in Section III-B, our model has two groups of free variables. One group is α_1, α_2, and μ which simulate the sustained temporal mechanisms and control the smooth effect of tube, and the other is β which controls the weight of temporal distortion variations. While ignoring μ as threshold and setting α to 0, the tube structure in our temporal model vanishes, that is, $D_{k,l,t} = d_{k,l,t}$ in (1). While setting β to 0, it is equivalent to that the final spatiotemporal distortion of a tube discards the temporal distortion, which is $\overline{D_t} = D_t$ in (5). In this experiment, we fix the μ at value 2.5 because we found the value of 2.5 can yield very good performance under different test conditions, and we focus on the evaluation of the impact of other three parameters on the distortion estimation performance. In the next, we give different combinations of α_1, α_2, β values to see the changes of performance. It should be noted that we use the WS-PSNR version of our method (i.e., OV-PSNR[WS-PSNR]) here since the WS-PSNR version showed the best performance in the previous performance comparison.

In the first row of Table IV, we validate the effectiveness of temporal distortion D'_t. The prediction accuracy of PLCC is greatly improved from 0.683 to 0.830, once we added the temporal distortion item into the calculation (set β from 0.0 to 1.0). This demonstrates the importance of temporal variation modeling in evaluating the omnidirectional video quality. The second row of Table IV shows a series of tests on different values of α_1, α_2, we found using the combination of 0.8 and 0.5 for α_1 and α_2 gives the best performance with fixing β to 1.0. In the last part of Table IV, we give more weight to temporal distortion (set β greater than 1.0), and the result reports a more inspiring performance. A possible explanation for this is that we may still underestimate the temporal distortion variations and need to build up a more sophisticated and precise spatiotemporal modeling approach.

TABLE IV
COMPARISON OF THE PERFORMANCES FOR DIFFERENT α_1, α_2, and β VALUES OF OV-PSNR[WS-PSNR] ON VR-VQA48.

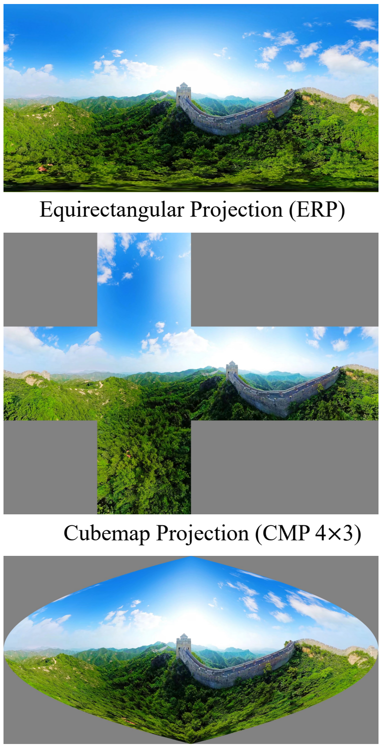
α_1, α_2	β	PLCC	SROCC	RMSE
0.0/0.0	0.0	0.683	0.663	9.040
0.0/0.0	1.0	**0.830**	**0.778**	**6.900**
0.2/0.1	1.0	0.832	0.780	6.866
0.5/0.25	1.0	0.835	0.782	6.805
0.6/0.5	1.0	0.838	0.788	6.764
0.7/0.5	1.0	0.838	0.789	6.751
0.8/0.5	1.0	**0.838**	**0.790**	**6.749**
0.9/0.5	1.0	0.838	0.787	6.764
0.95/0.5	1.0	0.836	0.788	6.778
0.8/0.5	0.3	0.805	0.758	7.348
0.8/0.5	3.0	0.851	0.814	6.493
0.8/0.5	10.0	**0.855**	**0.817**	**6.412**

D. Validation of Cross-format Quality Evaluation

In this subsection, we test the ability of cross-format quality evaluation of the proposed VQA method for omnidirectional video. More specifically, we validate the performance of our two OV-PSNR variants, i.e., OV-PSNR[S-PSNR] and OV-PSNR[CPP-PSNR]. Here we use the Lanzcos interpolation version of S-PSNR and CPP-PSNR for cross-format distortion measurement. As mentioned before, the original video source used in these two variants should be or be converted to the ERP format. Therefore, in this experiment, the original video format and impaired video format are chosen to ERP and CMP, respectively. These two formats are also the most commonly used in practice. However, the impaired videos in VR-VQA48 dataset are all ERP format, and we use the 360Lib toolset [15] to convert them to CMP 4×3 format. Fig. 8 shows an example of different projection formats of an omnidirectional video, where the test sequence Hangpai2 is used. The performance results of cross-format quality evaluation for 360 degree video is presented in Table V.

TABLE V
COMPARISON OF THE PERFORMANCES OF S-PSNR, CPP-PSNR AND OV-PSNR ACROSS PROJECTION FORMATS ON VR-VQA48.

Methods	PLCC	SROCC	RMSE
S-PSNR	0.551	0.563	10.329
CPP-PSNR	0.550	0.563	10.337
OV-PSNR[S-PSNR]	0.721	0.682	8.580
OV-PSNR[CPP-PSNR]	0.774	0.738	7.832


Equirectangular Projection (ERP)

Cubemap Projection (CMP 4×3)

Craster Parabolic Projection (CPP)

Fig. 8: An example of different types of projections for omnidirectional video.

As shown in Table V, we can see that on the basis of S-PSNR and CPP-PSNR, our corresponding methods (OV-PSNR[S-PSNR] and OV-PSNR[CPP-PSNR]) yield rather high video quality prediction performance. Here we use the same parameter settings as shown in the first performance experiment. Although our methods achieve considerable performance improvement, the two variants of OV-PSNR still do not reach the same good performance as the previous test using the same format VQA for omnidirectional video. The possible reason for this is that, our spatiotemporal model builds the tube structure only on ERP plane and reusing distortion from a different projection format for quality evaluation may cause a certain precision loss. Further work on the spatiotemporal modeling approach with constructing the tubes directly in spherical space may resolve this problem.

To summarize, we conduct experiments to validate the effectiveness of our spatiotemporal model extended to the existing quality metrics for omnidirectional video. The analysis results show that the performance of the existing quality metrics for omnidirectional video can be enhanced by our spatiotemporal model and the impact of temporal distortion variations is indeed important for omnidirectional video quality assessment.

V. CONCLUSION

In this paper, we have proposed a spatiotemporal modeling approach for evaluating the quality of omnidirectional video with consideration of both spatial and temporal characteristics of omnidirectional video. Specifically, we firstly construct a spatial-temporal tube-based structure as a basic quality assessment unit, to evaluate the average spatial distortion in temporal dimension at eye fixation level. Next, the smoothed distortion value of a tube is then consolidated by the temporal variations of distortion, which are calculated by the frequency and amplitude of the distortion gradient. Afterwards, the quality degradation score for the whole video sequence is obtained through an appropriate pooling method. Meanwhile, a full-reference objective VQA method has been presented, which can naturally integrate the three existing VQA metrics (S-PSNR, CPP-PSNR, WS-PSNR) for omnidirectional video into our spatiotemporal modeling approach. Finally, our experimentation validates the performance of our objective VQA method. The results show that our OV-PSNR provides a significant performance improvement compared to those quality metrics that depend only on spatial distortions.

In our spatiotemporal modeling approach, the block division and the construction of tube units are performed on planar space. Although significant quality assessment performance can be achieved by this proposed model, it may be more beneficial to consider the construction of the spatio-temporal tube directly in the spherical space. Further work may include this aspect. In addition, optimization of computational complexity for the procedure of tube construction may also be an interesting investigation direction. Moreover, our spatiotemporal model with a more sophisticated temporal pooling mechanism may achieve better performance.

REFERENCES

[1] C.-W. Fu, L. Wan, T.-T. Wong, and C.-S. Leung, "The rhombic dodecahedron map: An efficient scheme for encoding panoramic video," *IEEE Transactions on Multimedia*, vol. 11, no. 4, pp. 634–644, 2009.

[2] M. A. Usman, M. R. Usman, and S. Y. Shin, "A novel no-reference metric for estimating the impact of frame freezing artifacts on perceptual quality of streamed videos," *IEEE Transactions on Multimedia*, vol. 20, no. 9, pp. 2344–2359, 2018.

[3] Z. Wang, A. C. Bovik, H. R. Sheikh, E. P. Simoncelli *et al.*, "Image quality assessment: from error visibility to structural similarity," *IEEE transactions on Image Processing*, vol. 13, no. 4, pp. 600–612, 2004.

[4] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, vol. 2. Ieee, 2003, pp. 1398–1402.

[5] Z. Wang, L. Lu, and A. C. Bovik, "Video quality assessment based on structural distortion measurement," *Signal processing: Image communication*, vol. 19, no. 2, pp. 121–132, 2004.

[6] Z. Wang and Q. Li, "Video quality assessment using a statistical model of human visual speed perception," *JOSA A*, vol. 24, no. 12, pp. B61–B69, 2007.

[7] M. Yu, H. Lakshman, and B. Girod, "A framework to evaluate omnidirectional video coding schemes," in *2015 IEEE International Symposium on Mixed and Augmented Reality*. IEEE, 2015, pp. 31–36.

[8] V. Zakharchenko, K. P. Choi, and J. H. Park, "Quality metric for spherical panoramic video," in *Optics and Photonics for Information Processing X*, vol. 9970. International Society for Optics and Photonics, 2016, p. 99700C.

[9] Y. Sun, A. Lu, and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE signal processing letters*, vol. 24, no. 9, pp. 1408–1412, 2017.

[10] X. Xiu, Y. He, Y. Ye, and B. Vishwanath, "An evaluation framework for 360-degree video compression," in *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017, pp. 1–4.

[11] S. Yang, J. Zhao, T. Jiang, J. W. T. Rahim, B. Zhang, Z. Xu, and Z. Fei, "An objective assessment method based on multi-level factors for panoramic videos," in *2017 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2017, pp. 1–4.

[12] M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan, "Assessing visual quality of omnidirectional videos," *IEEE Transactions on Circuits and Systems for Video Technology*, 2018.

[13] M. H. Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on broadcasting*, vol. 50, no. 3, pp. 312–322, 2004.

[14] M. Narwaria, W. Lin, and A. Liu, "Low-complexity video quality assessment using temporal quality variations," *IEEE Transactions on Multimedia*, vol. 14, no. 3, pp. 525–535, 2012.

[15] J. Boyce, E. Alshina, A. Abbas, and Y. Ye, "JVET common test conditions and evaluation procedures for 360 video," *Joint Video Exploration Team of ITU-T SG*, vol. 16, 2017.

[16] S. Croci, C. Ozcinar, E. Zerman, J. Cabrera, and A. Smolic, "Voronoi-based objective quality metrics for omnidirectional video," in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2019, pp. 1–6.

[17] M. Orduna, C. Díaz, L. Muñoz, P. Pérez, I. Benito, and N. García, "Video multimethod assessment fusion (vmaf) on 360vr contents," *IEEE Transactions on Consumer Electronics*, vol. 66, no. 1, pp. 22–31, 2019.

[18] S. Croci, C. Ozcinar, E. Zerman, S. Knorr, J. Cabrera, and A. Smolic, "Visual attention-aware quality estimation framework for omnidirectional video using spherical voronoi diagram," *Quality and User Experience*, vol. 5, no. 1, pp. 1–17, 2020.

[19] H.-T. Lim, H. G. Kim, and Y. M. Ra, "Vr iqa net: Deep virtual reality image quality assessment using adversarial learning," in *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2018, pp. 6737–6741.

[20] C. Li, M. Xu, L. Jiang, S. Zhang, and X. Tao, "Viewport proposal cnn for 360deg video quality assessment," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10 177–10 186.

[21] J. Lubin, "The use of psychophysical data and models in the analysis of display system performance," in *Digital images and human vision*. MIT Press, 1993, pp. 163–178.

[22] C. J. van den Branden Lambrecht, "A working spatio-temporal model of the human visual system for image restoration and quality assessment applications," in *1996 IEEE International Conference on Acoustics, Speech, and Signal Processing Conference Proceedings*, vol. 4. IEEE, 1996, pp. 2291–2294.

[23] C. J. Van den Branden Lambrecht and O. Verscheure, "Perceptual quality measure using a spatiotemporal model of the human visual system," in *Digital Video Compression: Algorithms and Technologies 1996*, vol. 2668. International Society for Optics and Photonics, 1996, pp. 450–461.

[24] K. Tan, M. Ghanbari, and D. E. Pearson, "An objective measurement tool for MPEG video quality," *Signal processing*, vol. 70, no. 3, pp. 279–294, 1998.

[25] C. J. van den Branden Lambrecht, D. M. Costantini, G. L. Sicuranza, and M. Kunt, "Quality assessment of motion rendition in video coding," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 9, no. 5, pp. 766–782, 1999.

[26] S. Winkler, "Perceptual distortion metric for digital color video," in *Human Vision and Electronic Imaging IV*, vol. 3644. International Society for Optics and Photonics, 1999, pp. 175–184.

[27] A. B. Watson, Q. J. Hu, and J. F. McGowan, "Digital video quality metric based on human vision," *Journal of Electronic imaging*, vol. 10, no. 1, pp. 20–30, 2001.

[28] M. Barkowsky, J. Bialkowski, B. Eskofier, R. Bitto, and A. Kaup, "Temporal trajectory aware video quality measure," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 266–279, 2009.

[29] K. Seshadrinathan and A. C. Bovik, "Motion tuned spatio-temporal quality assessment of natural videos," *IEEE transactions on image processing*, vol. 19, no. 2, pp. 335–350, 2009.

[30] ——, "A structural similarity metric for video based on motion models," in *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, vol. 1. IEEE, 2007, pp. I–869.

[31] M. A. Masry and S. S. Hemami, "A metric for continuous quality evaluation of compressed video with severe distortions," *Signal processing: Image communication*, vol. 19, no. 2, pp. 133–146, 2004.

[32] A. Ninassi, O. Le Meur, P. Le Callet, and D. Barba, "Considering temporal variations of spatial visual distortions in video quality assessment," *IEEE Journal of Selected Topics in Signal Processing*, vol. 3, no. 2, pp. 253–265, 2009.

[33] P. Zhang and P. Gao, "Quality assessment for omnidirectional video with consideration of temporal distortion variations," in *2019 IEEE Visual Communications and Image Processing (VCIP)*. IEEE, 2019, pp. 1–4.

[34] M. Yuen and H. R. Wu, "A survey of hybrid mc/dpcm/dct video coding distortions," *Signal processing*, vol. 70, no. 3, pp. 247–278, 1998.

[35] U. Engelke, H. Kaprykowsky, H.-J. Zepernick, and P. Ndjiki-Nya, "Visual attention in quality assessment," *IEEE Signal Processing Magazine*, vol. 28, no. 6, pp. 50–59, 2011.

[36] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision research*, vol. 49, no. 10, pp. 1295–1306, 2009.

[37] V. Mahadevan and N. Vasconcelos, "Spatiotemporal saliency in dynamic scenes," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 1, pp. 171–177, 2009.

[38] D. Ćulibrk, M. Mirković, V. Zlokolica, M. Pokrić, V. Crnojević, and D. Kukolj, "Salient motion features for video quality assessment," *IEEE Transactions on Image Processing*, vol. 20, no. 4, pp. 948–958, 2010.

[39] J. E. Hoffman, "Visual attention and eye movements," *Attention*, vol. 31, pp. 119–153, 1998.

[40] R. Li, B. Zeng, and M. L. Liou, "A new three-step search algorithm for block motion estimation," *IEEE transactions on circuits and systems for video technology*, vol. 4, no. 4, pp. 438–442, 1994.

[41] M. Masry, S. S. Hemami, and Y. Sermadevi, "A scalable wavelet-based video distortion metric and applications," *IEEE Transactions on circuits and systems for video technology*, vol. 16, no. 2, pp. 260–273, 2006.

[42] A. Nasrabadi, M. Shirsavar, A. Ebrahimi, and M. Ghanbari, "Investigating the PSNR calculation methods for video sequences with source and channel distortions," in *2014 IEEE International Symposium on Broadband Multimedia Systems and Broadcasting*. IEEE, 2014, pp. 1–4.

[43] Y. Ye, E. Alshina, and J. Boyce, "Algorithm descriptions of projection format conversion and video quality metrics in 360lib," *Joint Video Exploration Team of ITU-T SG*, vol. 16, 2017.

[44] J. L. Bentley, "Multidimensional binary search trees used for associative searching," *Communications of the ACM*, vol. 18, no. 9, pp. 509–517, 1975.

[45] VR-VQA48 database, with subjective test data for panoramic video quality and head tracking data. [Online]. Available: https://github.com/Archer-Tatsu/head-tracking

[46] T. K. Tan, R. Weerakkody, M. Mrak, N. Ramzan, V. Baroncini, J.-R. Ohm, and G. J. Sullivan, "Video quality evaluation methodology and verification testing of HEVC compression performance," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 26, no. 1, pp. 76–90, 2015.

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

[48] V. Q. E. Group *et al.*, "Final report from the video quality experts group on the validation of objective models of video quality assessment," in *VQEG meeting, Ottawa, Canada, March, 2000*, 2000.

[49] H. R. Sheikh and A. C. Bovik, "Image information and visual quality," *IEEE Transactions on image processing*, vol. 15, no. 2, pp. 430–444, 2006.

[50] L. Zhang, L. Zhang, X. Mou, and D. Zhang, "FSIM: A feature similarity index for image quality assessment," *IEEE transactions on Image Processing*, vol. 20, no. 8, pp. 2378–2386, 2011.

[51] W. Xue, L. Zhang, X. Mou, and A. C. Bovik, "Gradient magnitude similarity deviation: A highly efficient perceptual image quality index," *IEEE Transactions on Image Processing*, vol. 23, no. 2, pp. 684–695, 2013.

[52] VMAF–Video Multi-Method Assessment Fusion. [Online]. Available: https://github.com/Netflix/vmaf

[53] C. Li, M. Xu, X. Du, and Z. Wang, "Bridge the gap between VQA and human behavior on omnidirectional video: A large-scale dataset and a deep learning model," in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 932–940.

[54] VQA-ODV database, a large-scale dataset of omnidirectional video for visual quality assessment. [Online]. Available: https://github.com/Archer-Tatsu/VQA-ODV