EAI Endorsed Transactions
on Industrial Networks and Intelligent Systems
Research Article EAI.EU

Retina-based quality assessment of tile-coded 360-degree videos

Nguyen Viet Hung[1,4], Pham Ngoc Nam[2,*], Truong Cong Thang[3], Bui Duy Tien[4], Nguyen Huu Thanh[4], Truong Thu Huong[4]

[1]East Asia University of Technology, Vietnam
[2]VinUniversity, Vietnam
[3]The University of Aizu, Aizuwakamatsu, Japan
[4]Hanoi University of Science and Technology, Vietnam

Abstract

Nowadays, omnidirectional content, which delivers 360-degree views of scenes, is a significant aspect of Virtual Reality systems. While 360 video requires a lot of bandwidth, users only see visible tiles, therefore a large amount of bitrate can be saved without affecting the user's experience on the service. The fact leads to current video adaptation solutions to filter out superfluous parts and extraneous bandwidth. To form a good basis for these adaptations, it is necessary to understand human's video quality perception. In our research, we contribute to building an effective omnidirectional video database that can be applied to study the effects of the five zones of the human retina. We also design a new video quality assessment method to analyze the impacts of those zones of a 360 video according to the human retina. The proposed scheme is found to outperform 22 current objective quality measures by 11 to 31% in terms of the PCC parameter.

Received on 18 May 2022; accepted on 20 June 2022; published on 21 June 2022

Keywords: Subjective quality assessment, foveation feature, virtual reality, picture quality, and image processing.

doi:10.4108/eetinis.v9i32.1058

1. Introduction

Nowadays, virtual reality technology has become popular, so it is of great interest to scientists. Virtual reality (VR) systems use omnidirectional content, which includes 360-degree views of scenes, to provide viewers with immersive experiences. Omnidirectional content is typically consumed utilizing Head-Mounted Displays as opposed to standard information displayed on a flat-screen (HMDs). In addition, a user only sees a small portion of the content (called a viewport) that corresponds to the current viewing direction at any given time [1].

In fact, because 360 videos have such a high bitrate, managing limitted system resources while ensuring user satisfaction (or called QoE - Quality of Experience) is a major challenge in omnidirectional content delivery. Especially, in the future, such a immersive service is expected to be delivered over sixth generation cellular networks which require a more comprehensive and prompt capture of QoE [2].

For this goal, many encoding and delivery schemes have been proposed in the literature [3] -[9]. Tiling-based viewport-adaptive streaming is one of the most used ways for 360 video streaming that is receiving a lot of attention from both academic areas and industry due to its ability to effectively reduce network bandwidth. A 360 video is spatially divided into small sections called tiles, each of which is encoded into numerous copies of varying quality levels in tiling-based viewport-adaptive streaming.

In general, when choosing a tile version, the visible tiles (those that overlap the viewport) are delivered in high resolution, while the other tiles are delivered in poor quality. Because users only see the visible tiles, a large amount of bitrate can be saved without affecting the user's experience.

To support the tile-based viewport adaptive streaming in the VR system, findings in the weight of different zones of a 360 video are highly necessary. Therefore,

*Pham Ngoc Nam. Email: nam.pn@vinuni.edu.vn

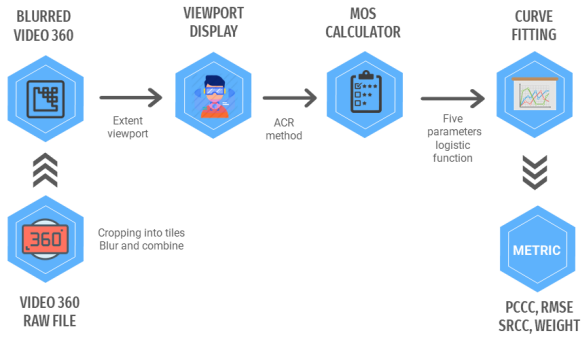


Figure 1. The proposed fitting model for impact of retina-related areas

the goals of our research is to study user perception of omnidirectional content, including the following contributions:

- To build a consistent subjective test scenario to accurately evaluate user perception on different retina-related zones of omnidirectional video. Thereby, we successfully established a database for omnidirectional videos that can be used for further research in the VR field.
- To build an efficient fitting method that can accurately find the impact weights of those different zones to user's quality experience on omnidirectional videos.

The overview of our overall analysis process is shown in Figure 1 which will be explained in detail in the following subsections. To the best of our knowledge, this is the first omnidirectional video database devoted to the effects of the five zones of the human retina. In this condition, we measure, for the first time, the effects of different zones on perceptual quality using a simple zone-weighted formulation. The zones corresponding to the fovea and parafovea of the human retina are discovered to be particularly significant for quality perception quantitatively.

The proposed fitting model is found to outperform 22 existing objective quality metrics under multicast video scenarios with heterogeneous quality, especially foveal quality index.

The following is the rest of the paper: Section 2 provides an overview of the state of the art. Section 3 describes our proposed method, which is followed by an experimental description in Section 4. Finally, Section 5 provides conclusions and future work.

2. Related Work

Recently, a wide variety of objective quality metrics have been proposed over several decades [10]- [28]. Some of these metrics take into account the foveation feature, hereafter referred to as foveal quality metrics, take into account the foveation feature [13]. All of these measures, however, are restricted to traditional content. So far, there has not been a foveal quality metric for omnidirectional material, except for the study of [28] mentioned, but we find the results still limited. Among of which, PSNR [26] is the most effective metric for assessing the quality of omnidirectional videos, according to experimental data. It is worth mentioning, however, that the photos utilized in that study are of consistent quality. PSNR is ineffective when the quality is spatially changeable. According to our survey, there has been no comprehensive study of objective quality metrics for omnidirectional images with tile-varying quality in the literature. In this paper, in terms of assessing the quality of omnidirectional videos, we will show that our proposed solution outperforms the 22 metrics: MSE, SSIM, MS-SSIM, UQI, VIFp, VIF, NQM, IW-PSNR, IW-SSIM, FSIM, FSIMc, SR-SIM, RFSIM, ADD-SSIM, PSIM, WSNR, FMSE, FPSNR, F-SSIM, GSIM, PSNR, ZWF by 11 to 31%.

On the other hand, subjective quality assessment also draws researchers' attention recently since omnidirectional images/videos become popular. There have been some research on subjective quality assessments of omnidirectional content [30]–[34]. In these researches, to generate images for user's rating in the subjective tests, numerous distortion types like as compression and Gaussian blur were used. In [30], the authors used 4 types of distortion including JPEG compression, JPEG2000 compression, Gaussian blur, and Gaussian noise. The authors in [31] only used one distortion type of H.265/HEVC compression. In the study [32], JPEG compression, JPEG2000 compression, and HEVC-intra compression were used. In [33], down sampling and JPEG compression were used to create the distorted images. In [34], video references are encoded using H265 encoding with a constant rate factor (CRF) = 10 and compressed with a quantization parameter (QP) = 22, 27, 32, 37, and 42 using the libx265 encoder of the FFmpeg tool. However, the above five studies only used images and videos with uniform distortion and did not contain images or videos of non-uniform quality. Therefore, these schemes are not suitable for VR video streaming, where user-focused areas should have high quality and less noticeable areas should have lower quality to save network resources. Furthermore, the foveation feature of the human eye is not taken into consideration when these databases are built.

Currently, there are also some studies on subjective quality assessments of images/videos with non-uniform quality in the literature [28], [35], [36], [37]. However, in the [36], [37]. the studies are only for traditional content without considering the omnidirectional contents. In [36], each image is split into four equal-width zones. To produce a distorted image with non-uniform quality,

after a fixed step size, the quality level of the zones are gradually decreased. According to the findings, when the step size is tiny, the difference in quality of experience (QoE) between non-uniform and uniform videos is negligible. Furthermore, the greatest step size that may be used without generating noticeable quality changes is determined by content characteristics. In [37], the authors divided each image into three zones with different quality levels: foveal, blending, and peripheral zones. The experimental results show that participants rarely perceive quality declines in the periphery zones with eccentricities greater than 7.5 degrees. Unlike [36] and [37], our research focuses on evaluating the quality of omnidirectional video experiences in five different regions. Therefore, it is possible to predict the quality deterioration at peripheral locations with an eccentricity higher than 7.5 degrees.

Besides, research [35] and [28] are the two researches to evaluate the quality of experience of omnidirectional content with non-uniform quality. The authors of [35] are interested in figuring out how to spatially lower image quality without affecting user perception. They propose that an omnidirectional image be divided into three zones corresponding to the macula, the near periphery, and the far periphery of the human retina. Each zone's image quality is gradually reduced until participants perceive a difference in perception. From the study results, a model including coding parameters for regions is proposed as a guide to reduce spatial image quality without loss of sensation. In addition, the authors have also shown that model-guided non-uniform quality scales accelerate image rendering by about 10× compared to the legacy schemes with uniform quality. However, the impacts of zones of the human retina on the quality of experience was not clearly quantified in [35]. Also, there is no performance evaluation of existing objective quality assessment metrics performed in [35]. In [28], from the original omnidirectional images, the authors extract the viewports to create the original dataset. In our study, the dataset generated from omnidirectional videos is more in line with the current trend when VR video streaming is of great interest. Based on zones of human retina, the authors in [28] divided each image into 5 zones respectively: the fovea, the parafovea, the perifovea, the near periphery, and the far periphery. To produce a distorted image with non-uniform quality, these zones will be encoded with different quality levels based on two scenarios: the quality degrades from the center to the periphery and vice versa. However, creating these non-uniform quality levels based on subjective factors is not suitable for real systems when factors on network conditions will mainly determine perceived quality.

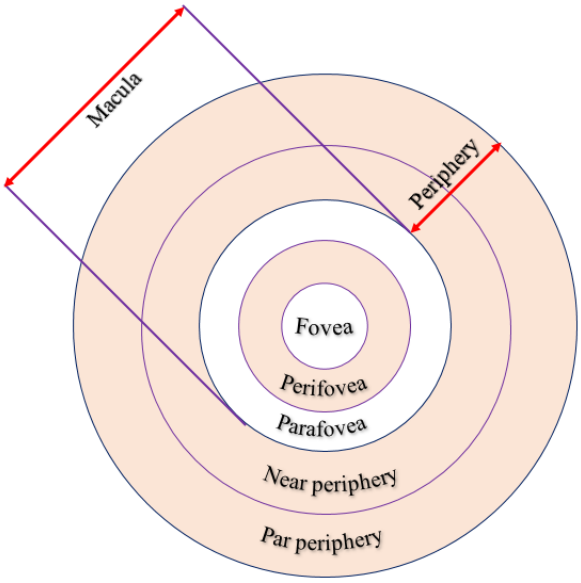


Figure 2. The retina divided into five regions.

Table 1. Zone eccentricity intervals

Zone	Z_1	Z_2	Z_2	Z_4	Z_5
Eccentricity interval (degrees)	0, 2.5	2.5, 4	4, 9	9, 30	30, $+\infty$

In this paper, we use tile-based non-uniform data set with quality levels of tiles being selected based on bandwidth traces, and head-movement traces. In addition, a measure to quantify the impact of regions on experience quality developed from mean squared error (MSE) is also presented. In our study, a new metric - WZUQI - was formed from the UQI index (universal image quality index). WZUQI was proven to outperform some common metrics such as PSNR and MSE under different types of image distortions.

3. Proposed solution to quantifying impacts of viewport zones to human's quality of experience

In this paper, we first propose a QoE metric for 360-video service, which is called $WZUQI$ (weighted-zone UQI). $WZUQI$ which will be used to investigate the effects of various zones on the perceived quality of 360° videos.

Second, we propose a new mapping function to predict QoE (or MOS) based on $WZUQI$ automatically, without requiring a large subject test measurement that should be done by a large pool of users. This mapping function is proven to predict the MOS effectively to real MOS rated by end users.

Let us elaborate our whole method step by step as follows:

3.1. Step 1: Formulate a new QoE metric - *WZUQI*

In our WZUQI method, a virtual viewport is divided into $K = 5$ zones $\{Z_k \mid 1 \leqslant k \leqslant K\}$ shown in Figure 2 with the corresponding zone eccentricities e shown in the TABLE 1. Zone Z_1 corresponds to the region of the fovea, zone Z_2 to the parafovea, zone Z_3 to the perifovea, zone Z_4 to the near periphery and zone Z_5 to the far periphery in the retina.

Let denote weight w_k $\{w_k \mid 1 \leqslant k \leqslant K\}$ that represents the impact of the Z_k region on the quality experience of human. Here w_k must satisfy the condition $\sum_{k=1}^{K} w_k = 1$.

Let $V = \{x_i \mid i = 1, 2, 3, ...N\}$ and $G = \{y_i \mid i = 1, 2, 3, ...N\}$ be the original viewport and the distorted viewport, respectively.

WZUQI is formed from the UQI index (universal image quality index) proposed by work [12]. This objective image quality index is mathematically determined with advantages such as: ease of computation, low computational complexity and independent of the images being tested, the viewing conditions or the individual observers. UQI was proven to outperform some common measures such as PSNR and MSE under different types of image distortions. The Universal Image Quality Index (UQI) is defined as follows:

$$UQI = \frac{1}{M}\left[\frac{4\sigma_{xy}\overline{x}\overline{y}}{(\sigma_x^2 + \sigma_y^2) \times [(\overline{x})^2 + (\overline{y})^2]}\right] \tag{1}$$

where:

M: the number of pixels of each image.
σ_x: loss of correlation.
σ_y: luminance distortion.
σ_{xy}: contrast distortion.
$\overline{x}$: is the average total of $\{x_i \mid i = 1, 2, 3, ..., N\}$
$\overline{y}$: is the average total of $\{y_i \mid i = 1, 2, 3, ..., N\}$

The parameter values are calculated as follows:

$$\overline{x} = \frac{1}{N}\sum_{i=1}^{N} x_i, \quad \overline{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})^2, \quad \sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \overline{y})^2$$

$$\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \overline{x})(y_i - \overline{y})$$

Then *WZUQI* is formed as follows:

$$WZUQI = \sum_{k=1}^{K} w_k UQI_k \tag{2}$$

3.2. Step 2 - Form a new MOS mapping function

After determining the *UQI* index of each region (UQI_k), the QoE objective metrics *WZUQI* is defined by formulae 2. Based on this QoE metric, we propose to form a new mapping function to calculate Mean Opinion Score (*MOS*). MOS is rated on a scale (1 – bad, 2 – poor, 3 – fair, 4 – good, and 5 – exceptional). Basically, *MOS* is given by users that reflects their satisfaction on a video service or any internet service. Initially the average MOS of a population must be calculated based on a large subjective test measurement. This subjective test require a lot of time to collect data as well as a large number of involved participants. Therefore, it is necessary to have a mapping function that can calculated a predicted MOS based on some QoE metric. And that predicted MOS should be highly correlated with the real MOS value rated by end users in reality.

In our paper, a 5-parameter logistic function is used to predict the MOS (Mean Opinion Score) values from *WZUQI* values. The 5-parameter logistic function has actually demonstrated a good performance in mapping between the objective quality indicators and MOS in the state of the art [26] and [38]. The formula to calculate the predicted MOS ($\widehat{MOS}$) can be computed as follows:

$$\widehat{MOS} = \alpha_1\left(\frac{1}{2} + \frac{1}{1 + e^{\alpha_2(WZUQI - \alpha_3)}}\right) + \alpha_4 WZUQI + \alpha_5 \tag{3}$$

where:

$\{\alpha_i \mid i = 1, 2, 3, 4, 5\}$ are the parameters to be fitted.

In our work, least squares fitting is used to fit the parameters α_i and the weights w_k, as described in [39]. In the following subsections, we will describe our testing scenario in order to evaluate the performance of *WZUQI* in terms of fitting accurately to the real MOS rated by real users. In our experiment, the proposed model runs on top of the tile-based non-uniform dataset formed by the research team to see how the performance would be affected. Figure 5 shows that the predicted MOS gets quite close to the subjective MOS rated by viewers.

4. Data Set Establishment

Our established dataset is formed from a subjective test measurement with 240 non-uniform viewports. Figure 3 shows the 95% confidence intervals of the MOS values. It can be shown that the scores cover the entire value range of 2.5 to nearly 3.5. At the two extremities of the score scale, the confidence intervals are typically smaller. This is because participants are more comfortable ranking very high (or poor) quality stimuli. In the following subsections, we will describe

Table 2. Features of the four 360 videos used in our experiments

VIDEOS	YOUTUBE ID	CONTENT	MOTION ACTIVITY
Diving	2OzlksZBTiA	Daytime diving, marine scene	Low
Paris	EkshFcLESPU	Sightseeing spots in Paris, daytime, tourist strolling	Low
RollerCoaster	8lsB-P8nGSM	RollerCoaster ride, outdoor, daytime	High
Venice	s-AJRFQuAtE	Virtual aerial view of buildings in Venice, in-outdoor, nightlight	Low

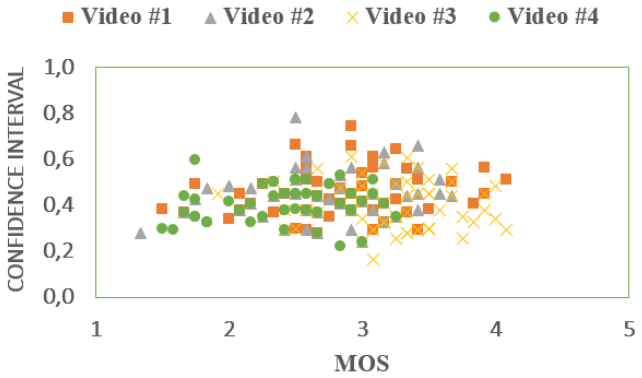


Figure 3. 95% confidence intervals of 240 MOS values.

how we prepared the videos and set up the subject test to achieve a reliable data set.

4.1. Video preparation

In this section, we will describe how this tile-based non-uniform data set is established and the experiment setup to capture consistent rating of users on the video quality. A data set with tighter confidence interval shows lower variability, which results in smaller margin of error. Therefore, designing a good experiment to achieve reliable data set that truly represents the large population of viewers is extremely important.

In our experiment, we use four 360° videos available on YouTube with different types of content such as indoors, under the ocean, natural landscapes, crowded streets, containing human face, day-light, night-light, etc. The purpose of covering a variety of different video textures is to achieve a data set that could represent a wide range of user quality experience. The specific characteristics of the four videos are described in the TABLE 2.

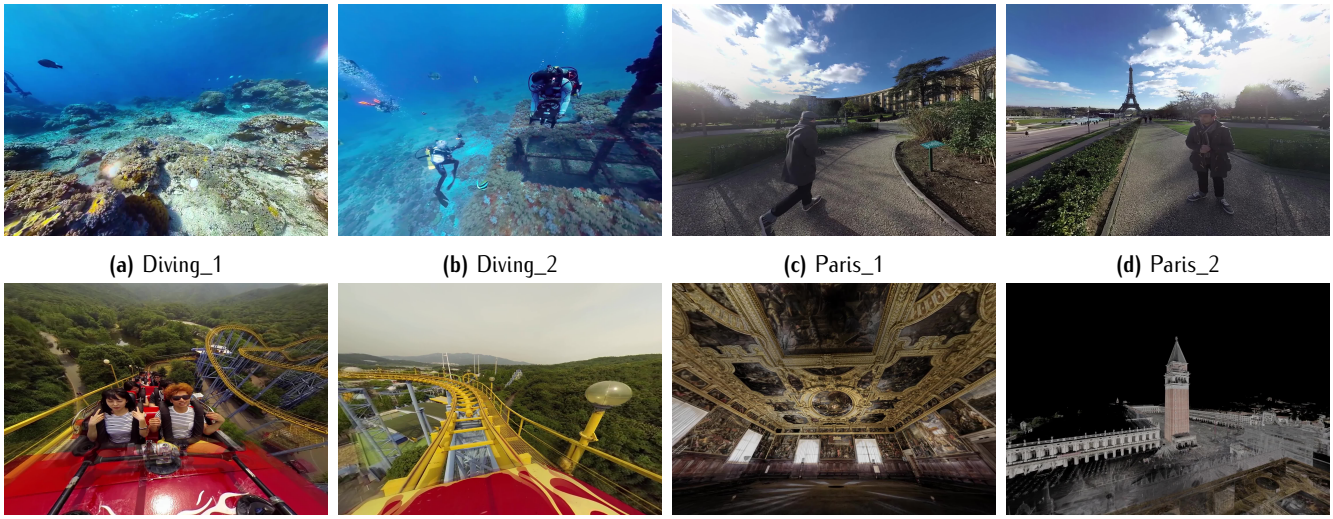
All these videos have a resolution of 3840 × 1920 (4K), 1792 frames, and a frame rate of 30 (fps). In order to create videos of non-uniform quality, the video is divided into $T = 64$ tiles (i.e., 8 × 8 tiling), each has a resolution of 480 × 240. We use the HEVC format to encode each tile into $N = 7$ versions corresponding to 7 QP values of 24, 28, 32, 36, 40, 44 and 48. From each streaming session, we select 60 viewports to ensure that the error values vary widely, corresponding to the user experience quality scores from high to low. Therefore, we finally get 240 tile-based images corresponding to 240 viewports, forming the desired tile-based dataset.

These viewports are collected continuously, in sequence of the context of the same video. In this way, the content-correlated viewports gives viewers a consistent quality experience like they are watching and assessing the quality of a video, rather than assessing sporadic viewports from different contexts or topics. With the dataset described above, we show some viewport samples extracted from four videos in Figure 4.

4.2. Subjective test set-up

To display the viewports, we use a set of devices consisting of a HTC VIVE PRO EYE headset and a computer. HTC VIVE PRO EYE has Dual-OLED displays with a combined resolution of 2880 × 1600 pixels and 615 PPI with the 110-degree field of view. In the testing, we employed the Absolute Category Rating approach [40], which was proved to be the best method in [37]. Before starting the testing process, participants are guided to familiarize themselves with the equipment, testing process and scoring. To avoid fatigue, the viewports from different videos are displayed alternately. Before an image is displayed, participants were asked to focus their gaze on a central point of the screen and hold that gaze as the image appears. For each viewport, participants spent 5 seconds maintaining their gaze, 10 seconds for rating the quality level, and then took a break of 5 seconds. Each participant gives a verbal score on a scale of 1 (poor) to 5 (excellent) that is then recorded by an assistant. The test is divided into 4 sessions, including 60 viewports/session with an average duration of no more than 20 minutes/session. The scores are collected from 40 participants with ages ranging from 18 to 40 years old. Following Recommendation ITU-T P.913 [41], a screening analysis of the collected scores is conducted. Consequently, three participants are rejected. The average scores of the valid participants are then used as MOSs of the corresponding images.

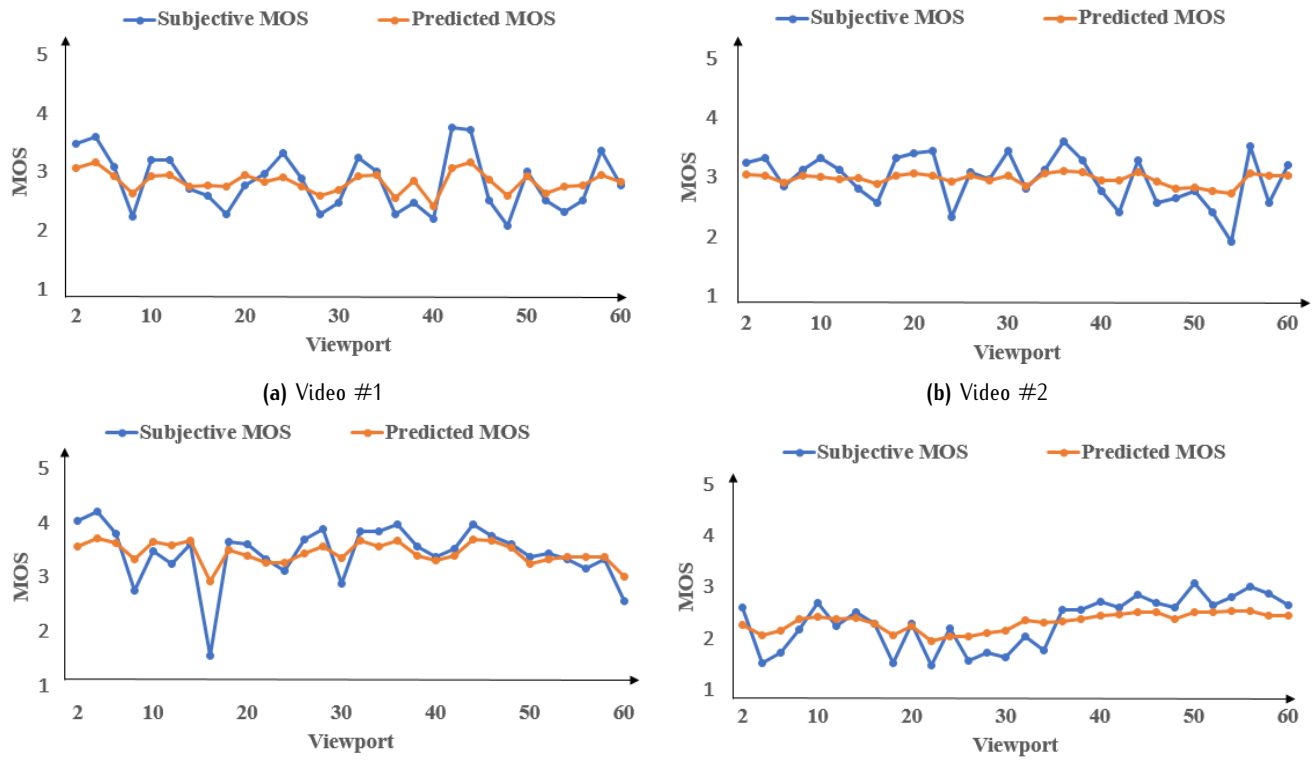
5. Results and Performance Evaluation



(a) Diving_1 (b) Diving_2 (c) Paris_1 (d) Paris_2

(e) Rollercoaster_1 (f) Rollercoaster_2 (g) Venice_1 (h) Venice_2

Figure 4. Some viewport samples are extracted from 4 videos



(a) Video #1 (b) Video #2

(c) Video #3 (d) Video #4

Figure 5. Predicted MOS vs. subjective MOS

5.1. Evaluation on quality assessment performance

In this section, we examine the performance of our proposed solution with 22 other existing objective quality metrics (OQM) with the same database scenario - our database. The notations and descriptions of the twenty-two metrics studied in this study are shown in TABLE 3. The purpose is to see if current measurements, particularly foveal quality metrics, are useful for assessing the quality of tile-varying omnidirectional videos. Except for the two metrics FPSNR and FSSIM which are implemented in work [23], [13] since their detailed implementation description is not available, the remaining metrics are provided by the

Table 3. An overview of the referenced methods

Metrics	Description
MSE [25]	The Mean Squared Error is calculated using visible pixels in a viewport with equal weights.
SSIM [10]	Structural SIMilarity was calculated using the structural similarity idea.
MS-SSIM [11]	Multi-scale SSIM is calculated using similar metrics obtained at various viewport resolutions (or multi-scales).
UQI [12]	Universal Image Quality, any distortion can be modeled as a combination of 3 factors: loss of correlation, luminance distortion, and contrast distortion.
VIFp [14]	The Visual Information Fidelity (VIFp) and the Visual Information Fidelity (VIF) in the pixel and wavelet domains were calculated using the relationships between picture information and visual quality.
VIF [14]	
NQM [15]	Noise Quality Measurement or the recovered distorted image's Signal-to-Noise Ratio in comparison to the model restored image.
IW-PSNR [16]	Combining information content weighting with PSNR metrics to create Information Content Weighted PSNR.
IW-SSIM [16]	Weighted SSIM for information content, which combines information content weighting with MS-SSIM measures.
FSIM [17]	Low-level feature weighting and local similarity measurements are combined to create feature similarity.
FSIMc [17]	Combining low-level feature weighting with local similarity metrics, feature similarity incorporates chromatic information.
SR-SIM [19]	Similarity based on spectral residuals, calculated using a spectral residual visual saliency model.
RFSIM [18]	Combining low-level feature weighting based on Riesz Transforms with local similarity measurements using Riesz Transforms.
ADD-SSIM [20]	The distribution of four metrics, including distortion position, distortion intensity, frequency changes, and histogram alteration, is considered in the analysis of distortion distribution.
PSIM [21]	The perceptual similarity combines the gradient magnitude similarities using two scales of color information similarity and a trustworthy perceptual-based pooling.
WSNR [22]	The weighted signal-to-noise ratio is the proportion of the average weighted signal power to the average weighted noise power, with the contrast sensitivity function as the weighting function.
FMSE [24]	The fovea, an area of the retina with the highest density of photoreceptors, has the highest visual acuity, with visual acuity rapidly decreasing for visual regions beyond the point of view.
FPSNR [23]	Foveal Peak Signal-to-Noise Ratio is calculated by combining PSNR measurements with weighting for each pixel based on the local frequency at that pixel.
F-SSIM [13]	Foveal-SSIM combines SSIM metrics with weighting for each macroblock based on the local frequency of pixels in that macroblock.
GSIM [27]	The omnidirectional photo's visual quality. A composite assessment of all weights and patch scores is used to estimate an image quality score.
PSNR [26]	Peak Signal-to-Noise Ratio (PSNR) is calculated by multiplying each pixel's weighting by the local frequency at that pixel.
ZWF [28]	This method, which focuses on the visual features of the human eye, was used to assess the quality of omnidirectional images and concentrate on various zones around the foveation point.

original author. In this study, to reflect what viewers actually watched, 22 metrics were calculated only for the viewports (i.e. visible pixels) of omnidirectional videos. We utilize the 360Lib software created by the Joint Video Experts Team (JVET) [42] to extract the viewports.

To quantify the fitting performance of the OQM metrics with MOS values, we used 3 performance metrics including: Pearson Correlation Coefficient (PCC), Root Mean Square Error (RMSE), and Spearman's ordered rank correlation coefficient (SROCC). Similar to our proposed metric, the OQM and MOS values are mapped using a five-parameter logistic function (i.e. (3)).

The PCC, SROCC, and RMSE values of the OQM metrics when fitting with all of the MOSs of the stimuli are shown in Figure 6 and the last columns of TABLES 4, 5, and 6. The PCC, SROCC and RMSE values for most measures change significantly between different source

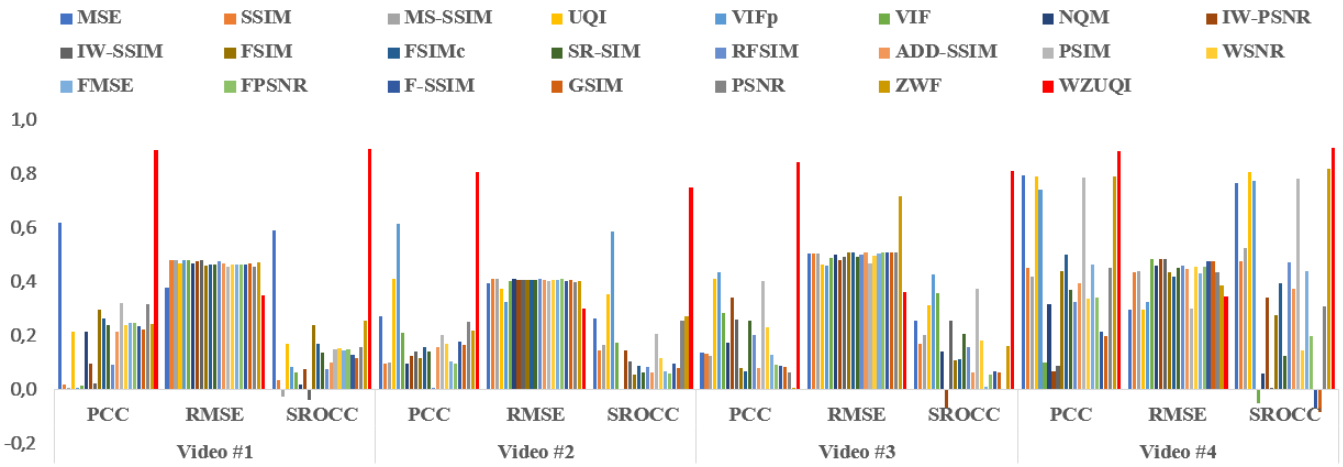


Figure 6. Our proposal vs. objective quality metrics

Table 4. SROCC values of indices were calculated with all stimuli and with stimuli of each source image between methods compared with the proposed method.

	Videos			
Metric	**Video #1**	**Video #2**	**Video #3**	**Video #4**
MSE [25]	0.591	0.262	0.253	0.767
SSIM [10]	0.034	0.146	0.170	0.476
MS-SSIM [11]	-0.026	0.165	0.200	0.524
UQI [12]	0.167	0.352	0.310	0.805
VIFp [14]	0.082	0.586	0.428	0.774
VIF [14]	0.061	0.174	0.358	-0.052
NQM [15]	0.017	0.005	0.141	0.057
IW-PSNR [16]	0.074	0.143	-0.070	0.339
IW-SSIM [16]	-0.039	0.103	0.253	0.004
FSIM [17]	0.238	0.055	0.106	0.276
FSIMc [17]	0.168	0.085	0.113	0.395
SR-SIM [19]	0.134	0.063	0.207	0.123
RFSIM [18]	0.075	0.085	0.154	0.472
ADD-SSIM [20]	0.100	0.063	0.064	0.375
PSIM [21]	0.147	0.206	0.375	0.782
WSNR [22]	0.152	0.117	0.179	0.142
FMSE [24]	0.144	0.067	0.010	0.438
FPSNR [23]	0.146	0.057	0.055	0.198
F-SSIM [13]	0.127	0.094	0.067	-0.070
GSIM [27]	0.115	0.077	0.062	-0.087
PSNR [26]	0.156	0.256	0.000	0.309
ZWF [28]	0.253	0.272	0.162	0.820
WZUQI	**0.894**	**0.750**	**0.809**	**0.895**

Table 5. PCC values of indices were calculated with all stimuli and with stimuli of each source image between methods compared with the proposed method.

	Videos			
Metric	**Video #1**	**Video #2**	**Video #3**	**Video #4**
MSE [25]	0.620	0.270	0.135	0.793
SSIM [10]	0.019	0.096	0.132	0.449
MS-SSIM [11]	0.002	0.098	0.125	0.419
UQI [12]	0.212	0.409	0.410	0.790
VIFp [14]	0.000	0.615	0.436	0.743
VIF [14]	0.012	0.210	0.282	0.101
NQM [15]	0.214	0.093	0.174	0.316
IW-PSNR [16]	0.096	0.121	0.340	0.066
IW-SSIM [16]	0.021	0.142	0.260	0.087
FSIM [17]	0.295	0.114	0.077	0.439
FSIMc [17]	0.262	0.154	0.069	0.500
SR-SIM [19]	0.239	0.140	0.253	0.370
RFSIM [18]	0.089	0.007	0.201	0.325
ADD-SSIM [20]	0.212	0.158	0.080	0.391
PSIM [21]	0.319	0.203	0.400	0.786
WSNR [22]	0.236	0.170	0.229	0.337
FMSE [24]	0.246	0.103	0.126	0.463
FPSNR [23]	0.245	0.095	0.092	0.340
F-SSIM [13]	0.232	0.177	0.087	0.212
GSIM [27]	0.221	0.166	0.083	0.199
PSNR [26]	0.318	0.251	0.063	0.450
ZWF [28]	0.244	0.217	0.000	0.791
WZUQI	**0.888**	**0.808**	**0.844**	**0.885**

videos. All of the metrics exhibit poor performance (PCC $\leqslant$ 0.79, SROCC $\leqslant$ 0.82, and RMSE $\geqslant$ 0.29). The PCC values of the foveal quality metrics (WSNR, FPSNR and F-SSIM) are also poor (i.e., from 0.087 to 0.339). This means that the current metrics are ineffective for analyzing the perceptual quality of omnidirectional videos with tile-varying quality.

Figure 5 shows two sets of values, respectively, including the subjective MOS obtained from the experiment and the predicted MOS obtained from the proposed quantitative method. We can observe that the two sets of MOS values have very identical trends. The predicted MOS set has a range of values at medium quality (i.e., from 2 to 4) and the difference is not too much between images. Meanwhile, the MOS set has a wider range of values and a larger difference in values between the images in the data set. On the other hand, in Figure 6, TABLE 4, 5, and 6, we can see that, the WZUQI formula outperforms the 22 existing OQM metrics for all source videos. The WZUQI formula has very high PCC, SROCC values and very low RMSE values. Specifically, the highest PCC and SROCC values, respectively, were 0.888 and 0.895 while the lowest RMSE values were 0.301. In particular, the WZUQI metric has the highest PCC and SROCC values for all four videos (Video #1, Video #2, Video #3, Video #4)

Table 6. RMSE values of indices were calculated with all stimuli and with stimuli of each source image between methods compared with the proposed method.

	Videos			
Metric	**Video #1**	**Video #2**	**Video #3**	**Video #4**
MSE [25]	0.375	0.395	0.505	0.295
SSIM [10]	0.478	0.408	0.505	0.433
MS-SSIM [11]	0.478	0.408	0.505	0.440
UQI [12]	0.467	0.374	0.464	0.297
VIFp [14]	0.478	0.323	0.458	0.325
VIF [14]	0.478	0.401	0.489	0.482
NQM [15]	0.467	0.408	0.502	0.460
IW-PSNR [16]	0.476	0.407	0.479	0.484
IW-SSIM [16]	0.478	0.406	0.492	0.483
FSIM [17]	0.457	0.407	0.508	0.436
FSIMc [17]	0.461	0.405	0.508	0.420
SR-SIM [19]	0.464	0.406	0.493	0.451
RFSIM [18]	0.476	0.410	0.499	0.459
ADD-SSIM [20]	0.467	0.405	0.508	0.446
PSIM [21]	0.453	0.401	0.467	0.299
WSNR [22]	0.465	0.404	0.496	0.457
FMSE [24]	0.463	0.408	0.505	0.430
FPSNR [23]	0.463	0.408	0.507	0.456
F-SSIM [13]	0.465	0.404	0.507	0.475
GSIM [27]	0.466	0.404	0.508	0.475
PSNR [26]	0.453	0.397	0.508	0.433
ZWF [28]	0.469	0.401	0.716	0.384
WZUQI	**0.348**	**0.301**	**0.362**	**0.344**

and the lowest RMSE values for three videos (Video #1, Video #2, Video #3). This result means that the WZUQI metric is rather effective for analyzing the perceptual quality of tile-based non-uniform dataset.

5.2. Impacts of the zones

As described in subsection 3, weights w_k are derived for each source video by the five-parameter logistic fitting function. Note that we use only the viewports of each video in our fitting. The values of the weights obtained through the experiment are shown in TABLE 7. As illustrated in TABLE 7, the w_k values of the four videos are approximately the same. This can be explained that during our subjective measurement experiment, users were asked to keep their gaze on the center of the screen. Except for w_1 and w_2, all other weighted values are low (i.e., $\leqslant$0.18). That shows that the zones with eccentricity $e \leqslant 4$ have little impact on human perception quality. Furthermore, the results of $w_1 \leqslant w_2 \leqslant w_3 \leqslant w_4 \leqslant w_5$ proves that near-center distortions have a more significant effect on user's quality experience than far-center distortions. In addition, the fovea region of the retina has the highest cone density, which explains why w_1 always has the highest value. Overall, the results show that although we use different video streams, our experimental results show that our method has good stability and is suitable for many different videos.

Table 7. Zone weights for each source video

	Video#1	*Video*#2	*Video*#3	*Video*#4	*Average*
w_1	0.4079	0.4097	0.4078	0.4073	0.4082
w_2	0.2612	0.2617	0.2614	0.2612	0.2614
w_3	0.1773	0.1766	0.1771	0.1774	0.1771
w_4	0.1108	0.1096	0.1108	0.1111	0.1105
w_5	0.0429	0.0424	0.0429	0.0430	0.0428

6. Conclusion

In this paper, we conducted subjective and objective quality assessments of omnidirectional video with tile-varying quality, with a focus on the human eye's foveation feature. We found that the sensitivity of the human eye and content scoring affect perceived quality. For perceptual quality, the zones of a viewport that correspond to the fovea and parafovea of the human eyes are particularly significant. With our experimental evaluations, we discovered that our scheme can improve PCC by 11% to 31% compared to the other 22 methods. In the future, we will extend contend genres and investigate quality oscillation patterns to gain insights into the perceptual habits and performance of the metrics.

Acknowledgement. This work was funded by Vingroup and supported by Vingroup Innovation Foundation (VINIF) under project code VINIF.2020.DA03.

References

[1] D. V. Nguyen, H. T. T. Tran, A. T. Pham, and T. C. Thang, "A new adaptation approach for viewport-adaptive 360-degree video streaming," in Proc. IEEE Int. Symp. Multimedia (ISM), Taichung, Taiwan, Dec. 2017, pp. 38–44.

[2] Taha, Abd-Elhamid M. 2021. "Quality of Experience in 6G Networks: Outlook and Challenges" Journal of Sensor and Actuator Networks 10, no. 1: 11. https://doi.org/10.3390/jsan10010011

[3] D. V. Nguyen, H. T. T. Tran, A. T. Pham and T. C. Thang, "An Optimal Tile-Based Approach for Viewport-Adaptive 360-Degree Video Streaming," in IEEE Journal on Emerging and Selected Topics in Circuits and Systems, vol. 9, no. 1, pp. 29-42, March 2019, doi: 10.1109/JETCAS.2019.2899488.

[4] J. Chakareski, R. Aksu, X. Corbillon, G. Simon, and V. Swaminathan,"Viewport-driven rate-distortion optimized 360◦ video streaming," in Proc. IEEE Int. Conf. Commun. (ICC), Kansas City, MO, USA, May 2018, pp. 1–7.

[5] C. Ozcinar, A. De Abreu, and A. Smolic, "Viewport-aware adaptive 360◦ video streaming using tiles for virtual reality," in Proc. IEEE Int. Conf. Image Process., Beijing, China, Sep. 2017, pp. 2174–2178.

[6] M. Hosseini and V. Swaminathan, "Adaptive 360 VR video streaming: Divide and conquer," in Proc. IEEE Int.

Symp. Multimedia (ISM), San Jose, CA, USA, Dec. 2016, pp. 107–110.
[7] M. Graf, C. Timmerer, and C. Mueller, “Towards bandwidth efficient adaptive streaming of omnidirectional video over HTTP: Design, implementation, and evaluation,” in Proc. 8th ACM Multimedia Syst. Conf. (MMSys), Taipei, Taiwan, Jun. 2017, pp. 261–271.
[8] A. Zare, A. Aminlou, M. M. Hannuksela, and M. Gabbouj, “HEVC compliant tile-based streaming of panoramic video for virtual reality applications,” in Proc. IEEE Picture Coding Symp. (PCS), Nuremberg, Germany, Dec. 2016, pp. 601–605.
[9] R. Skupin, Y. Sanchez, D. Podborski, C. Hellge, and T. Schierl, “HEVC tile based streaming to head mounted displays,” in Proc. 14th IEEE Annu. Consum. Commun. Netw. Conf. (CCNC), Las Vegas, NV, USA, Jan. 2017, pp. 613–615.
[10] Zhou Wang, A. C. Bovik, H. R. Sheikh and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," in IEEE Transactions on Image Processing, vol. 13, no. 4, pp. 600-612, April 2004, doi: 10.1109/TIP.2003.819861.
[11] Z. Wang, E. P. Simoncelli, and A. C. Bovik, "Multiscale structural similarity for image quality assessment," in Proc. 37th Asilomar Conf. Signals, Syst. Comput., vol. 2, Nov. 2003, pp. 1398–1402.
[12] Z. Wang and A. C. Bovik, “A universal image quality index,” IEEE Signal Process. Lett., vol. 9, no. 3, pp. 81–84, Mar. 2002.
[13] H. Ha, J. Park, S. Lee, and A. C. Bovik, “Perceptually unequal packet loss protection by weighting saliency and error propagation,” IEEE Trans. Circuits Syst. Video Technol., vol. 20, no. 9, pp. 1187–1199, Sep. 2010.
[14] H. R. Sheikh and A. C. Bovik, “Image information and visual quality,” IEEE Trans. Image Process., vol. 15, no. 2, pp. 430–444, Feb. 2006.
[15] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, “Image quality assessment based on a degradation model,” IEEE Trans. Image Process., vol. 9, no. 4, pp. 636–650, Apr. 2000.
[16] Z. Wang and Q. Li, “Information content weighting for perceptual image quality assessment,” IEEE Trans. Image Process., vol. 20, no. 5, pp. 1185–1198, May 2011.
[17] L. Zhang, L. Zhang, X. Mou, and D. Zhang, “FSIM: A feature similarity index for image quality assessment,” IEEE Trans. Image Process., vol. 20, no. 8, pp. 2378–2386, Aug. 2011.
[18] L. Zhang, D. Zhang, and X. Mou, “RFSIM: A feature based image quality assessment metric using Riesz transforms,” in Proc. 17th IEEE Int. Conf. Image Process., Sep. 2010, pp. 321–324.
[19] L. Zhang and H. Li, “SR-SIM: A fast and high performance IQA index based on spectral residual,” in Proc. 19th IEEE Int. Conf. Image Process.,Sep./Oct. 2012, pp. 1473–1476.
[20] K. Gu, S. Wang, G. Zhai, W. Lin, X. Yang, and W. Zhang, “Analysis of distortion distribution for pooling in image quality prediction,” IEEE Trans. Broadcast., vol. 62, no. 2, pp. 446–456, Jun. 2016.
[21] K. Gu, L. Li, H. Lu, X. Min, and W. Lin, “A fast reliable image quality predictor by fusing micro- and macro-structures,” IEEE Trans. Ind. Electron., vol. 64, no. 5, pp. 3903–3912, May 2017.
[22] N. Damera-Venkata, T. D. Kite, W. S. Geisler, B. L. Evans, and A. C. Bovik, “Image quality assessment based on a degradation model,” IEEE Trans. Image Process., vol. 9, no. 4, pp. 636–650, Apr. 2000.
[23] S. Lee and A. C. Bovik, “Foveated video image analysis and compression gain measurements,” in Proc. 4th IEEE Southwest Symp. Image Anal. Interpretation, Apr. 2000, pp. 63–67.
[24] A. Liu, W. Lin, and M. Narwaria, “Image Quality Assessment Based on Gradient Similarity,” IEEE Transactions on Image Processing, vol. 21, no. 4, pp. 1500-1512, April 2012.
[25] Rimac-Drlje, S., Vranješ, M. & Žagar, D. Foveated mean squared error—a novel video quality metric. Multimed Tools Appl 49, 425–445 (2010). https://doi.org/10.1007/s11042-009-0442-1
[26] H. T. T. Tran, C. T. Pham, N. P. Ngoc, A. T. Pham, and T. C. Thang, "A study on quality metrics for 360 video communications," IEICE Trans. Inf. Syst., vol. E101-D, no. 1, pp. 28–36, 2018.
[27] H. G. Kim, H. Lim and Y. M. Ro, "Deep Virtual Reality Image Quality Assessment With Human Perception Guider for Omnidirectional Image," in IEEE Transactions on Circuits and Systems for Video Technology, vol. 30, no. 4, pp. 917-928, April 2020, doi: 10.1109/TCSVT.2019.2898732.
[28] H. T. T. Tran, D. V. Nguyen, N. P. Ngoc, T. H. Hoang, T. T. Huong and T. C. Thang, "Impacts of Retina-Related Zones on Quality Perception of Omnidirectional Image," in IEEE Access, vol. 7, pp. 166997-167009, 2019, doi: 10.1109/ACCESS.2019.2953983.
[29] Vo, N. S., Lam, T. C., Bui, M. P., Phan, T. M., and Tran, Q. N. UAV Assisted Video Multicasting in 6G Networks: A Joint Caching and Trajectory Optimisation.
[30] H. Duan, G. Zhai, X. Min, Y. Zhu, Y. Fang, and X. Yang, "Perceptual quality assessment of omnidirectional images," in Proc. IEEE Int. Symp. Circuits Syst. (ISCAS), Florence, Italy, May 2018, pp. 1–5.
[31] M. Xu, C. Li, Z. Chen, Z. Wang, and Z. Guan, "Assessing visual quality of omnidirectional videos," IEEE Trans. Circuits Syst. Video Technol., to be published.
[32] H. G. Kim, H.-T. Lim, and Y. M. Ro, "Deep virtual reality image quality assessment with human perception guider for omnidirectional image," IEEE Trans. Circuits Syst. Video Technol., to be published.
[33] M. Huang, Q. Shen, Z. Ma, A. C. Bovik, P. Gupta, R. Zhou, and X. Cao, "Modeling the perceptual quality of immersive images rendered on head mounted displays: Resolution and compression," IEEE Trans. Image Process., vol. 27, no. 12, pp. 6039–6050, Dec. 2018.
[34] M. Elwardy, H. Zepernick and Y. Hu, "SSV360: A Dataset on Subjetive Quality Assessment of 360° Videos for Standing and Seated Viewing on an HMD," 2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW), 2022, pp. 01-06, doi: 10.1109/VRW55335.2022.00047.

[35] P. Guo, Q. Shen, Z. Ma, D. J. Brady, and Y. Wang, “Perceptual quality assessment of immersive images considering peripheral vision impact,” 2018, arXiv:1802.09065. [Online]. Available: https://arxiv.org/ abs/1802.09065

[36] P. Guo, Q. Shen, Z. Ma, D. J. Brady, and Y. Wang, Perceptual quality assessment of immersive images considering peripheral vision impact,” 2018, arXiv:1802.09065. [Online]. Available: https://arxiv.org/abs/1802.09065

[37] C.-F. Hsu, A. Chen, C.-H. Hsu, C.-Y. Huang, C.-L. Lei, and K.-T. Chen, "Is foveated rendering perceivable in virtual reality?: Exploring the efficiency and consistency of quality assessment methods," in Proc. 25th ACM Int. Conf. Multimedia, Mountain View, CA, USA, Oct. 2017, pp. 55–63.

[38] H. R. Sheikh, M. F. Sabir, and A. C. Bovik, "A statistical evaluation of recent full reference image quality assessment algorithms," IEEE Trans. Image Process., vol. 15, no. 11, pp. 3440–3451, Nov. 2006.

[39] Y.-F. Ou, Y. Xue, and Y. Wang, "Q-STAR: A perceptual video quality model considering impact of spatial, temporal, and amplitude resolutions," IEEE Trans. Image Process., vol. 23, no. 6, pp. 2473–2486, Jun. 2014.

[40] Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in Any Environment, document Rec. P.913 ITU-T, 2014.

[41] Recommendation ITU-T P.913, Methods for the Subjective Assessment of Video Quality, Audio Quality and Audiovisual Quality of Internet Video and Distribution Quality Television in Any Environment, 2014

[42] Joint Video Exploration Team. 360Lib. Accessed: 10th May 2022. [Online]. Available: https://www.itu.int/en/ITU-T/studygroups/2017-2020/16/Pages/video/prejvet.aspx/