A Viewport-adaptive Rate Control Approach for Omnidirectional Video Coding

Yichen Guo*, Mai Xu*, Li Yang*, and Rui Ding*

*School of Electronic and Information Engineering
Beihang University, Beijing, 100191, China
maixu@buaa.edu.cn (Corresponding author: Mai Xu)

Abstract

For omnidirectional videos (ODVs), the existing off-line coding approaches are designed based on the spatial or perceptual distortion in a whole ODV frame, ignoring the fact that subjects can only access viewports. To improve the subjective quality inside the viewports, this paper proposes an off-line viewport-adaptive rate control (RC) approach for ODVs in high efficiency video coding (HEVC) framework. Specifically, we predict the viewport candidates with importance weights and develop a viewport saliency detection model. Then, the predicted candidates and detected saliency are taken into account in our viewport-adaptive CTU traversal and bit allocation scheme. Finally, the experimental results validate that our approach is effective in saving bit-rates and improving subjective quality for encoding ODVs; meanwhile, our approach is also effective in the auxiliary task of saliency detection in viewports.

1 Introduction

With the rapid development of virtual reality (VR) applications, ODVs have gradually come into people's daily life. ODVs provide immersive and interactive viewing experience with 360° scenes, but requiring extremely high resolutions. Different from 2D videos, ODVs are normally watched through a head-mounted display (HMD), resulting that only the scene inside the viewport is visible. Therefore, there is a desirable demand on improving the coding efficiency of ODVs by considering the visible viewports. In contrast, the traditional video coding standards waste a huge amount of bits in representing and encoding a whole ODV frame.

During the past years, several approaches have been proposed to enhance coding efficiency of ODVs [1–11]. Considering the spatial stretching of equirectangular projection (ERP) format, some approaches [1,2,4,5] were proposed to optimize the spatial distortion by allocating bits according to projection-friendly peak signal to noise ratio (PSNR), such as spherical PSNR (S-PSNR) [12] or weighted to spherically uniform PSNR (WS-PSNR) [13]. However, these approaches do not take into account the perceptual quality of coded ODVs. Taking perception factors into consideration, some ODV coding approaches [4,6,7] were developed to combine the distribution of saliency into RC process or sampling density modification. Recently, there have emerged some on-line ODV coding approaches [8, 9, 11], in which the more bits are allocated to the subsequent viewport of the subject, which is predicted by algorithms or detected by the HMD. This ensures the higher fidelity of the viewports reached by subjects, as the regions outside the viewports cannot be accessed. Unfortunately, these approaches

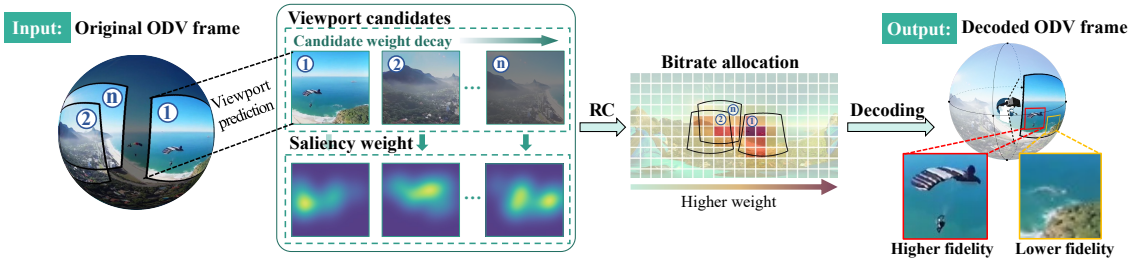

Figure 1: Brief framework of the proposed viewport-adaptive RC approach.

cannot be applied for off-line ODV coding, as there exist more than one potential viewports for different subjects.

To handle the above problems, we propose a viewport-adaptive ODV coding approach for off-line coding scenarios. As shown in Fig. 1, our scheme predicts the viewport candidates with decaying probability of being seen by subjects, and then detects the saliency within the viewport candidates. Given both the viewport candidates with decaying importance weights and the corresponding saliency, we develop the viewport-adaptive block traversal and bit allocation approach, which can effectively improve the ODV subjective quality in widely-used 2D video coding standards, such as HEVC [14]. In summary, the main contributions of this paper are as follows:

- We propose a viewport-adaptive ODV coding approach considering viewport candidates and their saliency, which effectively improves the subjective quality and saves the bit-rates.
- We develop a novel viewport saliency detection model as an auxiliary task of ODV encoding, which outperforms other state-of-the-art saliency detection models.

2 Related work

In recent years, a great amount of works have emerged for ODV coding. In the following, we review the existing approaches of ODV coding in three categories: spatial distortion optimization, perceptual quality promotion and viewport adaptation.

Spatial distortion optimization. In 2D video coding, the R-λ scheme [15] is a state-of-the-art RC approach. Considering the spatial distortion introduced by the ERP format, some ODV coding approaches have been proposed to modify the R-λ scheme for bit allocation according to the weight maps of projection-friendly PSNR. For example, Liu et al. [4] proposed a novel RC approach for ODV coding, which optimizes S-PSNR with sphere-based mean square error (S-MSE) as signal distortion. Li et al. [5] proposed assigning bits in the CTUs according to the weight map of WS-PSNR. Moreover, in [1], both representation and encoding are maintained in the spherical domain, such that the signal distortion by the sphere-to-plane projection can be reduced. However, these approaches do not consider the perceptual quality of subjects, which are determined by their viewports.

Perceptual quality promotion. Recently, several works have been proposed to optimize perceptual quality of ODV coding by considering the salient regions in the whole ODV frame. As a pilot study, Sitzmann et al. [6] subjectively validated the effectiveness of saliency-based sampling scheme on ODVs, in which the saliency map is generated by ground-truth head movement data. To improve perceptual quality,

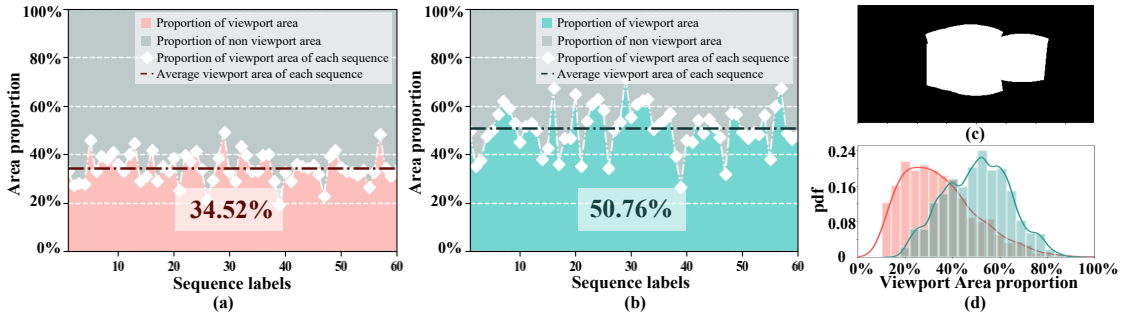

Figure 2: (a) Average percentage of viewport regions over all the subjects for each ODV sequence; (b) Average percentage of viewport regions over all the frames for each ODV sequence; (c) Viewport mask with white regions representing viewport; (d) The frequency distribution of viewport area proportion over all subjects and all ODV frames.

Luz et al. [7] proposed a saliency-based encoding solution and a relevant perceptual quality metric, both of which are driven by their saliency prediction model. In addition, Liu et al. [4] proposed determining bit allocation of ODV coding by optimizing perceptual distortion upon the front-center-bias of human attention. Although the above approaches consider salient regions in ODV coding, they ignore that only the viewports ($\sim 12.5\%$ of a whole ODV frame) of ODVs can be viewed by subjects.

Viewport adaptation. Considering that subjects can only see the content inside the viewports, several on-line ODV coding approaches have been proposed to allocate more bits to the viewport regions. Nasrabadi et al. [8] proposed using layered encoding approach, in which the base layer covers the whole frame and the enhancement layers only refer to viewports detected by the HMD. As such, the perceptual quality can be improved through interaction between subjects and the HMD. From the perspective of compression-friendly projection, Kuzyakov et al. [11] proposed encoding ODVs with a pyramid geometry. The base of the pyramid is the full resolution viewport, while the sides of the pyramid gradually decrease in quality for other regions. Unfortunately, these approaches are all on-line coding, requiring the subsequent viewport of each subject, and they cannot be applied to off-line ODV coding, as there are more than one potential viewport for different subjects.

3 The proposed approach

3.1 Motivation

Here, we thoroughly analyze the VQA-ODV dataset [16] to explore human behavior on viewing ODVs. First, we follow [12] to project the ODV viewports into 2D plane, and obtain the pixel-wise viewport mask in ERP format, as shown in Fig. 2-(c). Next, we integrate viewport areas in all ODV frames for each individual subject, and calculate the proportion of integrated viewports to the whole frame. The proportion is then averaged over all subjects. For each ODV sequence, we plot the proportion of viewport and non-viewport areas in Fig. 2-(a). Similarly, we calculate the proportion of integrated viewports by all subjects in the same frame, and then averaged over all frames (See in Fig. 2-(b)). From Fig. 2-(a) and (b), we can see that the averaged proportions of viewport areas are 34.52% and 50.76%, respectively. This implies that

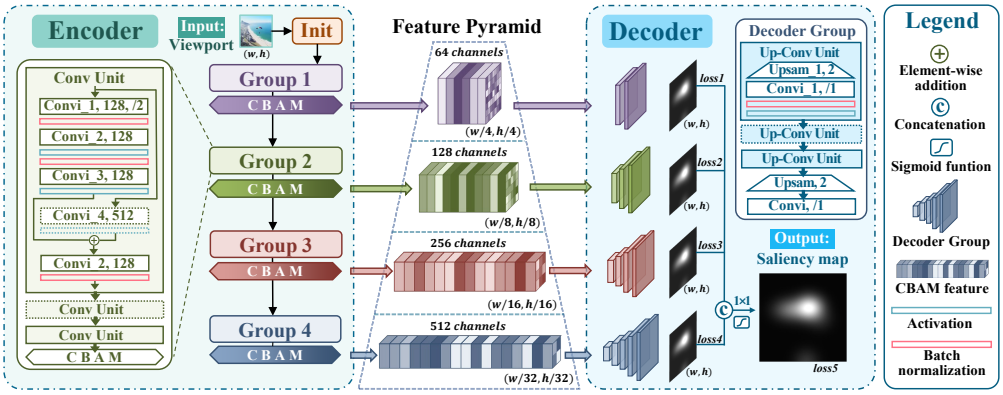

Figure 3: The architecture of viewport saliency detection.

there exists a large amount of perceptual redundancy in each ODV frame. Furthermore, Fig. 2-(d) shows the frequency distribution of viewport area proportion over all subjects and all ODV frames, respectively. Similar results of perceptual redundancy can be also found from this figure. Above analyses motivate us to propose the offline viewport-adaptive RC approach for ODVs to simultaneously save bit-rates and improve the subjective quality. Our approach is composed of three stages. For stage I, the viewport candidates are predicted with importance weights. For stage II, the viewport saliency is detected for each viewport candidate. For stage III, the viewport-adaptive block traversal and bit allocation are developed for subjective quality based ODV coding. More details are presented as follows.

3.2 Stage I: Viewport candidate prediction

In our approach, we adopt our previous work of the viewport proposal network (VP-Net) [16] for predicting viewport candidates on ODVs. Here, we briefly review the VP-Net. Each ODV frame and its optical flow are fed into a spherical convolutional neural network (CNN) [17], which can avoid the space-varying distortion introduced by planar projection. Subsequently, the extracted feature maps are fused and fed into two parallel 1×1 convolutional layers to separately generate the viewport offsets and importance weights. Finally, the viewports are selected after the viewport softer non maximum suppression (NMS) to flow into the viewport saliency detection model.

3.3 Stage II: Viewport saliency detection

In this stage, we develop a novel fully convolutional network (FCN) for viewport saliency detection, which is shown in Fig. 3. As can be seen, we incorporate multi-scale and multi-level saliency information into an encoder-decoder architecture, to predict pixel-wise saliency map. Between the encoder and decoder, the attention feature pyramid is proposed to cover different sizes of receptive fields. In the following, we concretely introduce the encoder, feature pyramid and decoder.

• **Encoder.** The predicted viewport from stage I is first fed into the initial group for feature extraction, which includes a convolution and max-pooling layer. Then, the feature flows into four residual groups for splitting multi-level spatial features, which are based on Resnet-34 [18]. Furthermore, an attention module is proposed to

embed between two adjacent residual groups for adaptive feature refinement, which is inspired by Convolutional Block Attention Module (CBAM) [19].

• **Feature Pyramid.** After encoder, multi-level features are hierarchically collected in a buffer, named feature pyramid. The feature pyramid fully exploits low-, middle- and high-level information, which predicts human attention with various receptive fields. Specifically, in each tier of the pyramid, the feature maps contain increasing channel numbers (i.e., 64, 128, 256 and 512) with decreasing map sizes (i.e., $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{16}$ and $\frac{1}{32}$). Then, the multi-level features in the feature pyramid are fed into the decoder for feature re-scaling and information fusion.

• **Decoder.** For the decoder, it up-samples the multi-level features from feature pyramid, to restore the map size and fuse multi-level saliency estimation. Specifically, the feature from each pyramid tier is fed into the corresponding decoder group, which is comprised by multiple "Up-Conv" units. Here, we propose Up-Conv unit to restore the size of feature map, which contains an up-sampling layer and a convolution layer. Instead of using deconvolution, our Up-Conv units can overcome the uneven overlap in feature maps. As shown in Fig. 3, each decoder group contains increasing number of Up-Conv units (i.e., 2, 3, 4 and 5), to meet with multi-scale features from the feature pyramid. Following Up-Conv units, the estimated saliency maps are concatenated, and then fed into a 1×1 convolution with *softmax* non-linearity to generate the final saliency map.

• **Loss function of model training.** The loss function for training each viewport is based on the losses of saliency maps at m levels from different decoder groups, as follows:

$$\mathcal{L} = \lambda_1 \cdot \mathcal{L}_{\mathrm{KL}} + \lambda_2 \cdot \mathcal{L}_{\mathrm{CC}} = \lambda_1 \sum_{m=1}^{M} D_{\mathrm{KL}}(\mathbf{S}_m || \hat{\mathbf{S}}_m) + \lambda_2 \sum_{m=1}^{M} (1 - R_{\mathrm{CC}}(\mathbf{S}_m, \hat{\mathbf{S}}_m)), \quad (1)$$

where $\hat{\mathbf{S}}_m$ and $\mathbf{S}_m$ denote the m-th level predicted saliency map and the corresponding ground-truth saliency map, respectively. Here, $\mathcal{L}$ includes two parts of loss, i.e., $\mathcal{L}_{\mathrm{KL}}$ and $\mathcal{L}_{\mathrm{CC}}$, which measure the Kullback-Leibler (KL) divergence and linear correlation coefficient(CC) between $\hat{\mathbf{S}}_m$ and $\mathbf{S}_m$, respectively. Moreover, $D_{\mathrm{KL}}(\cdot)$ and $R_{\mathrm{CC}}(\cdot)$ denote the calculation operations of KL and CC, respectively. Additionally, λ_1 and λ_2 are two trade-off hyper-parameters balancing the loss between $\mathcal{L}_{\mathrm{KL}}$ and $\mathcal{L}_{\mathrm{CC}}$.

3.4 Stage III: Viewport-adaptive rate control

In this section, we introduce our viewport-adaptive RC approach, which includes viewport-adaptive CTU traversal and bit allocation. Given the set of L predicted viewport candidates $\{\mathbf{V}_l\}_{l=1}^{L}$, corresponding weights $\{w_l\}_{l=1}^{L}$ at stage I and detected viewport saliency maps $\{\mathbf{S}_l\}_{l=1}^{L}$ at stage II, our viewport-adaptive RC approach is implemented as follows.

First, we propose a viewport-adaptive CTU traversal strategy based on the viewport candidates and its saliency maps, for further bit allocation on CTU level. In the CTU traversal, we assign the j-th pixel of the i-th CTU a weight $w_{i,j}$ in different manners, depending on whether the pixel is inside the viewport candidates or not.

Algorithm 1: Viewport-adaptive ODV Rate Control.

Input: RC parameters c_i and k_i, target bits R, sets of predicted viewport candidates $\{\mathbf{V}_l\}_{l=1}^{L}$ and corresponding weights $\{w_l\}_{l=1}^{L}$ and detected viewport saliency maps $\{\mathbf{S}_l\}_{l=1}^{L}$.
Output: Bit allocation r_i for each CTU.

$w_{i,j} \leftarrow 0$, $a_i \leftarrow c_i k_i$, $b_i \leftarrow \frac{1}{k_i+1}$
for Each CTU $\mathbf{B}_i$ **do**
  if $\mathbf{B}_i \cap \{\mathbf{V}_l\}_{l=1}^{L} \neq \varnothing$ **then**
    Calculate $w_{i,j}$ using [16], $w_i \leftarrow \frac{1}{J}\sum_{j=1}^{J} w_{i,j}$
  end
  else
    $w_i \leftarrow \varepsilon$ ($\varepsilon \rightarrow 0^+$)
  end
  Calculate λ in (9) with w_i using RTE method.
  Calculate corresponding r_i using (8).
  return r_i for each CTU.
end

For pixels inside the viewport candidates, we combine the candidate weight and the viewport saliency according to [16]. Otherwise, we set to it to a positive value close to zero.

Given the weight of each pixel $w_{i,j}$, the weight of the i-th CTU w_i is calculated as the average weight of all pixels in it:

$$w_i = \frac{1}{J}\sum_{j=1}^{J} w_{i,j}. \tag{2}$$

Next, given a fixed bit-rate R, the goal of bit allocation for ODVs is to minimize the overall perceptual distortion d_i inside the viewports within a frame F_t. The formulation can be expressed by

$$\min \sum_{i=1}^{I} d_i, \quad \text{s.t.} \sum_{i=1}^{I} r_i = R, \tag{3}$$

where d_i and r_i are the distortion and target bits of each CTU, respectively. We use the hyperbolic model [20] to formulate the relationship between distortion and bit-rate as

$$d_i = c_i r_i^{-k_i}, \tag{4}$$

where c_i and k_i are parameters related to ODV content. The values of c_i and k_i are updated after coding each ODV frame, and then used for the next ODV frame with the same frame level. According to the CTU weight assignment illustrated above, the viewport-adaptive distortion can be minimized as follows:

$$\min \frac{\sum_{i=1}^{I}(w_i \cdot d_i)}{\sum_{i=1}^{I} w_i} \quad \text{s.t.} \sum_{i=1}^{I} r_i = R. \tag{5}$$

According to the R-λ scheme, (5) can be converted into an unconstrained optimization problem:

$$\min G, \text{ where } G = \left(\frac{\sum_{i=1}^{I} w_i \cdot d_i}{\sum_{i=1}^{I} w_i}\right) + \lambda \sum_{i=1}^{I} r_i. \tag{6}$$

The solution to (6) can be obtained by setting its derivative to zero, which is

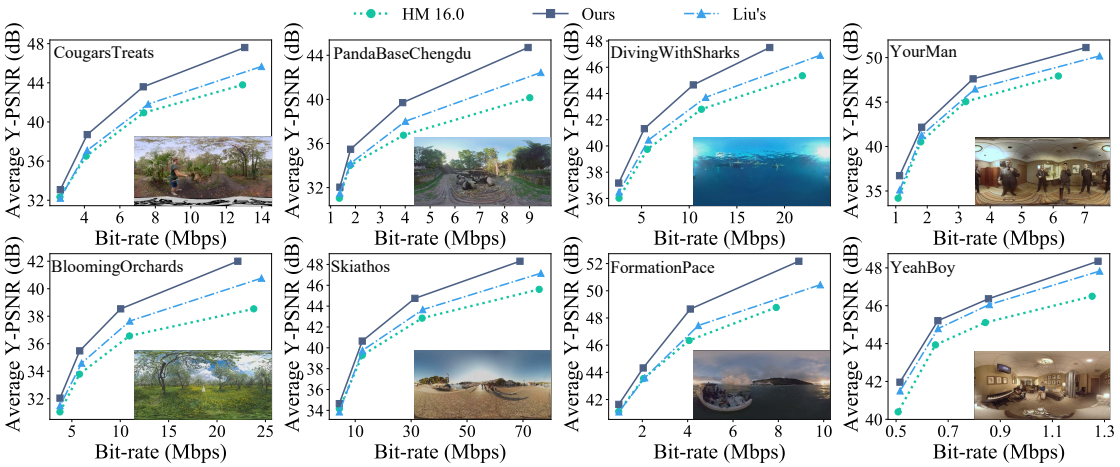

Figure 4: Rate-distortion comparison among our approach, Liu's approach and the R-λ anchor.

$$\lambda = -(\frac{w_i}{\sum_{i=1}^{I} w_i})\frac{\partial d_i}{\partial r_i} = (\frac{w_i}{\sum_{i=1}^{I} w_i}) \cdot c_i k_i r_i^{-k_i-1}. \tag{7}$$

Assume that $a_i = c_i k_i$, $b = \frac{1}{k_i+1}$ and $\widetilde{w}_i = \frac{w_i}{\sum_{i=1}^{I} w_i}$. Then, the allocated bits for the i-th CTU $\mathbf{B}_i$ is formulated as

$$r_i = (\frac{\widetilde{w}_i a_i}{\lambda})^{b_i}. \tag{8}$$

Given the target bits R of one frame, the following holds,

$$\sum_{i=1}^{I} r_i = \sum_{i=1}^{I} (\frac{\widetilde{w}_i a_i}{\lambda})^{b_i} = R. \tag{9}$$

Finally, we can get the r_i for each CTU $\mathbf{B}_i$ by solving (9), using the recursive Taylor expansion (RTE) solution [21]. The whole procedure of our viewport-adaptive RC approach is summarized in Algorithm 1.

4 Experiment

4.1 Settings

In this section, experiments are conducted to validate the performance of our viewport-adaptive RC approach and viewport saliency detection. Our approach was implemented on HEVC test model (HM) 16.0, and its default R-λ scheme is utilized as an anchor. Furthermore, we compared our approach with Liu's approach [4], which is also an R-λ based RC approach for ODV coding. Our experiments used the standard test sequences of AVS [22] in raw format with resolution of 8K (7680×3840 pixels). They are all at 29 fps in duration from $10 \sim 20$ seconds, with contents of indoor and outdoor scenes including people, animals and so on. The target bits were set to be the same as four fixed quality parameters (QPs) of 27, 32, 37 and 42. We used the eye-tracking weighted PSNR (EWPSNR) [23] and information content weighted structural similarity (IW-SSIM) [24] as the quality metric by using the ground-truth viewports and eye fixation data in VQA-ODV dataset.

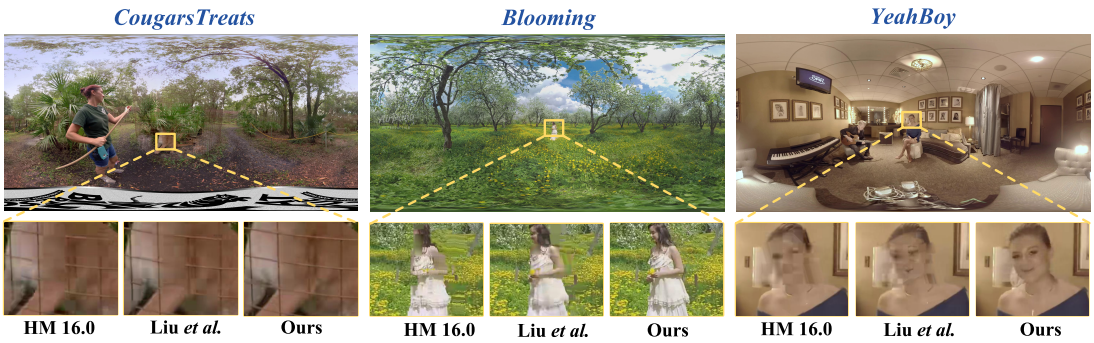

Figure 5: Examples of visual quality comparison among our approach, Liu's approach and the R-λ anchor approach.

4.2 Evaluation on coding performance

RD performance. We compare the rate-distortion (R-D) performance of our, Liu's and the default R-λ approaches using EWPSNR in Y channel. We plot in Fig. 4 the R-D curves of all test sequences. It can be seen that our scheme achieves higher EWPSNR than both Liu's approach and and the standard encoder at the same target bit-rates, for all test sequences.

BD-BR and BD-quality. We also test the Bjøntegaard delta bit rate (BD-BR) savings according to the improvement of Bjøntegaard delta PSNR (BD-PSNR) and Bjøntegaard delta SSIM (BD-SSIM) for all test sequences, respectively. As shown in Table. 1, the maximum of BD-PSNR and BD-SSIM improvement with corresponding BD-BR savings are up to 2.39 dB and -35.61%; 3.41 $\times 10^{-2}$ and -56.38%, respectively. The average BD-PSNR and BD-SSIM improvement with corresponding BD-BR savings among all test sequence are 1.91 dB and -26.09%; 1.12 $\times 10^{-2}$ and -29.55%, respectively. It can be seen that our approach performs much better than Liu's approach in terms of BD-PSNR, BD-SSIM and the corresponding BD-rate.

Table 1: BD-rate saving and BD-PSNR improvement for each test sequence.

Sequences	BD-PSNR(dB)		BD-rate(%)		BD-SSIM($\times 10^{-2}$)		BD-rate(%)	
	Ours	Liu's	Ours	Liu's	Ours	Liu's	Ours	Liu's
CougarsTreats	2.39	0.59	-22.92	-6.88	1.80	0.82	-20.46	-14.01
PandaBaseChengdu	2.73	1.01	-35.61	-15.79	3.41	2.19	-56.38	-28.10
DivingWithSharks	1.98	0.70	-33.50	-13.80	0.02	0.08	-30.27	-16.08
YourMan	1.97	0.79	-19.79	-9.06	1.03	0.98	-22.26	-19.97
BloomingOrchards	1.72	1.04	-22.66	-17.17	0.71	0.93	-10.22	-4.56
Skiathos	1.68	0.57	-29.57	-12.06	1.54	1.42	-44.36	-12.38
FormationPace	1.53	0.87	-26.14	-20.33	0.36	0.23	-36.53	-20.87
YeahBoy	1.29	0.89	-18.52	-11.89	0.05	0.05	-16.00	-3.28
Average	1.91	0.81	-26.09	-13.37	1.12	0.84	-29.55	-14.91

Subjective quality. Furthermore, Fig. 5 shows the visual quality of three randomly selected test ODVs, encoded by HM-16.0 with ours, Liu's and standard R-λ scheme at the same target bit-rates, respectively. It can be obviously seen that our approach yields much better visual quality than the other two approaches, with smaller blurring effect and less artifacts. Moreover, our approach effectively alleviates the blocking effect, which can severely impact the visual quality, as shown in Fig. 5, at regions such as the handrail edge, human body, human face and so on. In summary, our approach outperforms both Liu's approach and the standard R-λ scheme in RD performance, BD-quality, BD-BR and subjective quality.

4.3 Evaluation on Saliency Detection

Here, we evaluate the effectiveness of our saliency detection model. The performance is compared with four other state-of-the-art saliency detection model, via four metrics: KL divergence, CC, normalized scanpath saliency (NSS) and the area under the receiver operating characteristic curve (AUC). The larger values of CC, NSS or AUC indicate higher accuracy in saliency detection, while a smaller KL divergence means better performance of saliency detection. Table 2 tabulates the results of KL divergence, CC, NSS and AUC for our own and four other models over ground-truth viewports of AVS test sequence. The best and the second-best results are highlighted in red and blue, respectively. We can see from this table that our model outperforms other compared approaches, with at least 0.053 increase in CC, 0.079 decrease in KL divergence and 0.234 increase in NSS. This experiment verifies the effectiveness of our model in saliency detection on ground-truth viewports.

Table 2: Comparison of saliency prediction results among ours and other compared approaches.

Metrics	Ours	SAM [25]	IRL [26]	BMS [27]	SaltiNet [28]
KL divergence ↓	0.963 ± 0.381	3.215 ± 2.202	1.042 ± 0.402	1.356 ± 0.454	1.832 ± 1.185
CC ↑	0.577 ± 0.182	0.524 ± 0.180	0.509 ± 0.170	0.295 ± 0.167	0.294 ± 0.166
NSS ↑	2.039 ± 1.604	1.805 ± 1.526	1.518 ± 1.145	1.268 ± 1.151	1.440 ± 0.830
AUC ↑	0.884 ± 0.187	0.884 ± 0.184	0.843 ± 0.188	0.802 ± 0.194	0.875 ± 0.068

5 Conclusion

In this paper, we have proposed an off-line viewport-adaptive RC approach for ODV coding. According to our analysis of ground-truth head movement, there exists large amount of perceptual redundancy in ODV coding. For this reason, we proposed to separately deal with the CTUs inside and outside the viewports. Specifically, our approach predicts the viewport candidates with importance weights for stage I. Then, we developed a novel deep learning model of viewport saliency detection for stage II. Based on the above two stages, we proposed a CTU traversal approach and a weight-based RC formulation for stage III. Next, the RTE solution was presented to solve the weight-based RC formulation with minimal perceptual distortion at each encoded ODV frame. Finally, the experimental results showed the effectiveness of our approach in both bit-rate saving and perceptual quality improvement for ODV coding over the HEVC framework and in the auxiliary task of viewport saliency detection.

References

[1] Yiming Li, Jizheng Xu, and Zhenzhong Chen, "Spherical domain rate-distortion optimization for omnidirectional video coding," *IEEE TCSVT*, 2018.

[2] Minhao Tang, Yu Zhang, Jiangtao Wen, and Shiqiang Yang, "Optimized video coding for omnidirectional videos," 2017.

[3] Madhukar Budagavi, John Furton, Guoxin Jin, and Ankur Saxena, "360 degrees video coding using region adaptive smoothing," in *IEEE ICIP*, 2015.

[4] Yufan Liu, Li Yang, Mai Xu, and Zulin Wang, "Rate control schemes for panoramic video coding," *Journal of Visual Communication and Image Representation*, 2018.

[5] BiJia Li, Li Song, Rong Xie, and WenJun Zhang, "Weight-based bit allocation scheme for vr videos in hevc," in *IEEE VCIP*, 2017.

[6] Vincent Sitzmann, Ana Serrano, Amy Pavel, Maneesh Agrawala, Diego Gutierrez, Belen Masia, and Gordon Wetzstein, "Saliency in vr: How do people explore virtual environments?," *IEEE transactions on visualization and computer graphics*, 2018.

[7] Guilherme Luz, João Ascenso, Catarina Brites, and Fernando Pereira, "Saliency-driven omnidirectional imaging adaptive coding: Modeling and assessment," in *MMSP*, 2017.

[8] Afshin TaghaviNasrabadi, Anahita Mahzari, Joseph D Beshay, and Ravi Prakash, "Adaptive 360-degree video streaming using layered video coding," in *IEEE VR*, 2017.

[9] Duc V Nguyen, Huyen TT Tran, Anh T Pham, and Truong Cong Thang, "An optimal tile-based approach for viewport-adaptive 360-degree video streaming," *IEEE Journal on Emerging and Selected Topics in Circuits and Systems*, 2019.

[10] Liyang Sun, Fanyi Duanmu, and Liu, "A two-tier system for on-demand streaming of 360 degree video over dynamic networks," *IEEE JESTCS*, 2019.

[11] E. Kuzyakov and D. Pio., "Next-generation video encoding techniques for 360 video and vr. [online].," 2016.

[12] Matt Yu, Haricharan Lakshman, and Bernd Girod, "A framework to evaluate omnidirectional video coding schemes," in *IEEE ISMAR*, 2015.

[13] Yule Sun, Ang Lu, and Lu Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE signal processing letters*, 2017.

[14] Gary J Sullivan, Jens-Rainer Ohm, Woo-Jin Han, and Thomas Wiegand, "Overview of the high efficiency video coding (hevc) standard," *IEEE TCSVT*, 2012.

[15] Bin Li, Houqiang Li, Li Li, and Jinlei Zhang, "λ domain rate control algorithm for high efficiency video coding," *IEEE TIP*, 2014.

[16] Mai Xu, Lai Jiang, Chen Li, Zulin Wang, and Xiaoming Tao, "Viewport-based cnn: A multi-task approach for assessing 360° video quality," *IEEE TPAMI*, 2020.

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

[18] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *IEEE ICCV*, 2016.

[19] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon, "Cbam: Convolutional block attention module," in *ECCV*, 2018.

[20] Min Dai, Dmitri Loguinov, and Hayder Radha, "Rate-distortion modeling of scalable video coders," in *IEEE ICIP*, 2004.

[21] Shengxi Li, Mai Xu, Zulin Wang, and Xiaoyan Sun, "Optimal bit allocation for ctu level rate control in hevc," *IEEE TCSVT*, 2016.

[22] IEEE1857.9 1st Beijing. 2016. 1857.9-01-n0001 output document., ," .

[23] Zhicheng Li, Shiyin Qin, and Laurent Itti, "Visual attention guided bit allocation in video compression," *Image and Vision Computing*, 2011.

[24] Zhou Wang and Qiang Li, "Information content weighting for perceptual image quality assessment," *IEEE TIP*, 2010.

[25] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara, "Predicting human eye fixations via an lstm-based saliency attentive model," *IEEE TIP*, 2018.

[26] Chen Xia, Junwei Han, Fei Qi, and Guangming Shi, "Predicting human saccadic scanpaths based on iterative representation learning," *IEEE TIP*, 2019.

[27] Jianming Zhang and Stan Sclaroff, "Exploiting surroundedness for saliency detection: a boolean map approach," *IEEE TPAMI*, 2015.

[28] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor, "Scanpath and saliency prediction on 360 degree images," *Signal Processing: Image Communication*, 2018.