Viewport Proposal CNN for 360° Video Quality Assessment

Chen Li†, Mai Xu†‡*, Lai Jiang†, Shanyi Zhang† and Xiaoming Tao§
† School of Electronic and Information Engineering, Beihang University, Beijing, China
‡ Hangzhou Innovation Institute (HZII), Beihang University, Hangzhou, Zhejiang, China
§ Department of Electronic Engineering, Tsinghua University, Beijing, China
†{jnlichen123, maixu, jianglai.china, shanyi_zhang}@buaa.edu.cn

Abstract

Recent years have witnessed the growing interest in visual quality assessment (VQA) for 360° video. Unfortunately, the existing VQA approaches do not consider the facts that: 1) Observers only see viewports of 360° video, rather than patches or whole 360° frames. 2) Within the viewport, only salient regions can be perceived by observers with high resolution. Thus, this paper proposes a viewport-based convolutional neural network (V-CNN) approach for VQA on 360° video, considering both auxiliary tasks of viewport proposal and viewport saliency prediction. Our V-CNN approach is composed of two stages, i.e., viewport proposal and VQA. In the first stage, the viewport proposal network (VP-net) is developed to yield several potential viewports, seen as the first auxiliary task. In the second stage, a viewport quality network (VQ-net) is designed to rate the VQA score for each proposed viewport, in which the saliency map of the viewport is predicted and then utilized in VQA score rating. Consequently, another auxiliary task of viewport saliency prediction can be achieved. More importantly, the main task of VQA on 360° video can be accomplished via integrating the VQA scores of all viewports. The experiments validate the effectiveness of our V-CNN approach in significantly advancing the state-of-the-art performance of VQA on 360° video. In addition, our approach achieves comparable performance in two auxiliary tasks. The code of our V-CNN approach is available at `https://github.com/Archer-Tatsu/V-CNN`.

1. Introduction

Nowadays, 360° video has become part of our daily life. Typically, people watch 360° video through a head-mounted display (HMD), resulting in that only the content within the viewport is visible. They can obtain immersive experience by changing their viewports in the range of

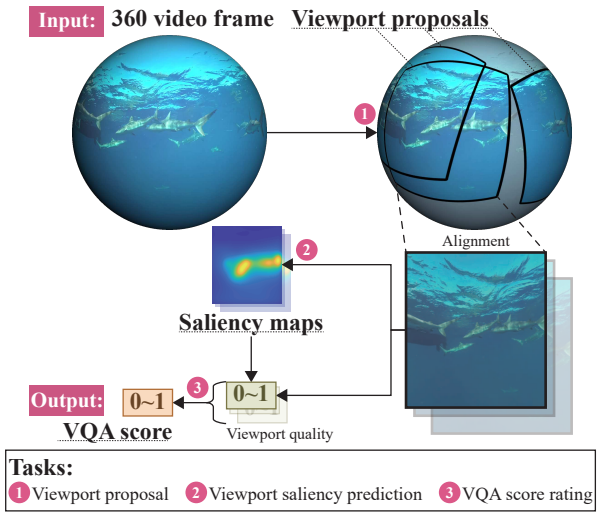


Figure 1. Brief framework of the proposed V-CNN for 360° video quality assessment with multiple tasks.

$360 \times 180°$. Therefore, 360° video covers the $360 \times 180°$ content, requiring extremely high resolution for clear representation. To deliver high-resolution 360° video over the bandwidth-limited channel, compression is urgently needed [4, 37] to save encoding bit-rates, meanwhile causing dramatic degradation on visual quality. Moreover, the quality degradation in the viewport can be more noticeable, since people only watch a small part of the original frame. It is thus necessary to study objective visual quality assessment (VQA) for 360° video, which can be used as the guidance for processing 360° video.

Recently, there has been increasing interest in assessing the visual quality of 360° image/video [19, 21, 38, 39, 32, 28]. For example, based on peak signal-to-noise ratio (PSNR), Yu *et al.* [38] proposed sphere-based PSNR (S-PSNR) to calculate PSNR by considering uniform sampling on sphere. They also provide an alternative VQA approach that calculates PSNR in the ground-truth viewports; however, it is not practical as the viewport is not available when

*Mai Xu is the corresponding author of this paper.

processing 360° video offline. Li *et al.* [19] proposed a model with convolutional neural networks (CNNs), which directly utilizes the data of head and eye movement (HM and EM) in VQA on 360° video. Based on the HM and EM data, the patches are extracted for assessing visual quality of 360° video. Unfortunately, the existing practical VQA approaches for 360° video ignore the fact that the subjects can only access the viewports, rather than the patches or original frames. Thus, a viewport-based VQA approach is more reasonable to reflect the subjective quality of 360° video. Note that when 360° video is compressed offline or broadcasted, the viewports of multiple subjects need to be predicted.

The procedure of watching 360° video can be divided into two stages. In the first stage, the subjects control the HM to select the attractive viewport visible to them, while the region outside the viewport is invisible. Thus, the position of viewport is determined by the HM of the subjects. In the second stage, the subjects may control the EM to focus on a small region (normally 2-3 visual degrees) of the viewport, i.e., salient region that is clearly captured by eyes. Due to this, it is reasonable to assess the visual quality of 360° video by taking both potential viewport (HM) and saliency within viewport (EM) into consideration. Recall that the region-based CNN (R-CNN) approaches achieve the great success in object detection [9, 24, 12], with the designs of region proposal structure. Inspired by these key factors, it is intuitive to design a viewport-based CNN approach with viewport proposal for VQA on 360° video.

In this paper, we propose a viewport-based CNN (V-CNN) approach for VQA on 360° video, the framework of which is briefly shown in Figure 1. This approach consists of two stages, i.e., viewport proposal and VQA. In the stage of viewport proposal, several potential viewports are proposed for the next stage, as the result of predicting HM. In the stage of VQA, after viewport alignment, the saliency map and VQA score are generated for each proposed viewport. At last, the final VQA score is averaged over VQA scores of all viewport proposals. As is shown in Figure 1, there are three tasks in our V-CNN approach. (1) Viewport proposal: This task is to propose potential viewports containing attractive content for subjects. (2) Saliency prediction in viewport: This task is to predict saliency map of each proposed viewport. (3) VQA score rating: This task is to rate the VQA score of the 360° video based on each proposed viewport with the predicted saliency map. Although the main task is VQA, our V-CNN approach is able to obtain comparable results in other two auxiliary tasks, i.e., viewport proposal and viewport saliency prediction.

The main contributions of this paper are: (1) We propose a viewport-based approach for VQA on 360° video[1], on the basis of real human behaviors when watching 360° video. (2) We embed auxiliary tasks to boost VQA, by predicting the HM and EM of subjects. (3) Our approach outperforms the state-of-the-art approaches in VQA, and achieves comparable performance in the auxiliary tasks, i.e., viewport proposal and viewport saliency prediction.

[1]The code of our V-CNN approach is available at `https://github.com/Archer-Tatsu/V-CNN`.

2. Related Works

VQA on 360° video. In addition to the traditional VQA approaches for 2D video, the spherical characteristic has been taken into account in the latest VQA works for 360° video [28, 32, 38, 39], which advance the metric of PSNR. For example, considering the nonuniformity of mapping spherical content to plane, both weighted-to-spherically-uniform PSNR (WS-PSNR) [28] and area weighted spherical PSNR (AW-SPSNR) [32] apply weight allocation in calculating PSNR to balance the nonuniform sampling density. In addition, there are some other VQA approaches [38, 39] aiming to solve the nonuniform sampling in 360° video. In [38], Yu *et al.* proposed S-PSNR that resamples the original frame with a set of uniformly distributed points on sphere. In [39], Zakharchenko *et al.* proposed applying Craster parabolic projection (CPP) to calculate PSNR, named CPP-PSNR. The effectiveness of these approaches [28, 32, 38, 39] has been verified through VQA experiments on different 360° video datasets [30, 29, 22]. Recently, there have also emerged several deep learning based approaches [19, 21] for VQA on 360° video. Li *et al.* [19] proposed utilizing both HM and EM data in a CNN model for VQA on 360° video. Lim *et al.* [21] proposed a 360° video VQA approach with adversarial learning. Both [19] and [21] apply the same patch-based approach [17] as VQA on 2D video, and learned to sample and allocate weights on patches for VQA on 360° video.

Saliency models on 360° video. Recently, there has been increasing interest in studying attention models for 360° video. Several 360° video datasets [6, 31, 7] were built, including HM data [6, 31] and even EM data [7]. These datasets enable further research of analysis and even prediction on human attention. Based on the datasets, the human attention of watching 360° video was thoroughly studied [23, 26], by investigating consistency and fixation bias among subjects. In addition, several works [34, 35, 3] were proposed to predict the HM and EM of subjects when watching 360° video. For predicting HM, Cheng *et al.* [3] developed a spatial-temporal deep neural network (DNN) to predict saliency maps of HM on 360° video. In [3], the original 360° videos are remapped to cubemap projection, followed by cube padding before being fed to the network. Most recently, Xu *et al.* [34] have also proposed predicting viewports on 360° video, achieved by developing a new framework of deep reinforcement learning (DRL). For predicting EM, a DNN was developed in [35] to predict saliency maps of EM on 360° video. In addition,

some works [20, 1] focused on predicting the scan-paths of HM/EM in 360° image/video.

However, none of above VQA approaches for 360° video considers the fact that the subjective visual quality highly depends on the viewports in 360° video. Additionally, there is no existing VQA work for 360° video that benefits from embedding auxiliary tasks. Thus, in this paper, we propose a viewport-based approach for VQA on 360° video, which considers predicting the real human behaviors, i.e., HM and EM, when watching 360° video.

3. The Proposed V-CNN Approach

3.1. Framework

Since the visual quality of an impaired 360° frame only depends on the viewports seen by the subjects, it is intuitive to propose some potential viewports for VQA. In our V-CNN approach, there are two stages, i.e., viewport proposal (Stage I) and VQA (Stage II), for assessing full-reference (FR) VQA score of the t-th impaired sequence frame $\mathbf{F}_t$. The framework of the V-CNN approach is shown in Figure 2. As is seen in this figure, the input to V-CNN includes the t-th impaired frame $\mathbf{F}_t$, the t-th reference frame $\mathbf{F}_t^{\text{ref}}$ and the $(t-\Delta t)$-th impaired frame $\mathbf{F}_{t-\Delta t}$.

At stage I, the potential viewports on the impaired sequence frames are proposed, which also performs as the first auxiliary task of our V-CNN approach. Specifically, we propose a viewport proposal network (VP-net) to generate the viewport candidates and their corresponding importance weights. Then, a viewport softer non maximum suppression (NMS) is developed to filter the viewport candidates for proposal, according to the importance weights.

At stage II, the proposed viewports from stage I as well as their corresponding error maps are first converted from spherical signals to plane, through our developed viewport alignment method. Then, we develop a viewport quality network (VQ-net) to assess the VQA score of each viewport. In the VQ-net, the saliency map of each viewport is predicted by a mini-DenseNet, which also performs as the second auxiliary task of our V-CNN. Given the error map and the predicted saliency map, the VQA score of each viewport is rated by a shallow CNN structure. At last, the final VQA score of the impaired 360° frame is obtained, by averaging over all VQA scores of the proposed viewports. This accomplishes the main task of our V-CNN, i,e, FR VQA on impaired 360° frames. Note that we follow [38] to yield the final score of FR VQA on an impaired 360° video sequence, given the corresponding reference sequence.

3.2. Stage I: Viewport proposal

In this stage, we take impaired frame $\mathbf{F}_t$ and its temporal residual ($\mathbf{F}_t - \mathbf{F}_{t-\Delta t}$) as the inputs. Subsequently, the VP-net and viewport softer NMS are developed to propose potential viewports. These viewport proposals can be seen as the results of the first auxiliary task for modelling general HM of subjects on 360° video sequence. More details are presented as follows.

VP-net. The architecture of the VP-net is shown in Figure 3. Specifically, the inputs of the impaired frame and temporal residual are resampled with a widely used spherical grid [14], and then concatenated together to flow into the following layers. Since 360° video sequence is a spherical signal, we apply the spherical CNN [5] in our VP-net to extract spherical features. Note that there are two types of convolutional layers in spherical CNN, which implement convolution on the unit sphere, denoted as S^2, and 3D rotation group, denoted as SO(3), respectively. The outputs of both these two types of layers are feature maps on SO(3). Additionally, the outputs of the 1st, 3rd and 5th layers are down-sampled and then concatenated with the outputs of the 7th to 9th layers, in order to fuse the low-level information into the deep layers. After the last spherical convolutional layer, the spherical feature maps on the S^2 are obtained, denoted as $\mathbf{T}$, by being converted from SO(3).

Given the spherical features $\mathbf{T}$ on the S^2, we further apply the anchor implementation [24] on sphere for viewport proposal. For each pixel location in $\mathbf{T}$, there exists a corresponding anchor location on the sphere. This anchor location can be represented as $\mathbf{v}^{\text{a}} = (\phi^{\text{a}}, \theta^{\text{a}})$, where ϕ^{a} and θ^{a} are the longitude and latitude, respectively. Accordingly, the spherical feature maps $\mathbf{T}$ can be regarded as feature vectors at the anchor locations. Since the size of the viewport is fixed when subjects are watching 360° video, each viewport can be represented as $\mathbf{v} = (\phi, \theta)$, where ϕ and θ are the longitude and latitude of its center, respectively. To be more specific, for each of the I anchor locations $\{\mathbf{v}_i^{\text{a}}\}_{i=1}^{I}$, a viewport candidate is generated by predicting the offset $\Delta\mathbf{v}_i = (\Delta\phi_i, \Delta\theta_i)$ to $\mathbf{v}_i^{\text{a}}$ and its corresponding importance weight $\hat{w}_i$. This is achieved via two convolutional layers with kernel size of 1. Subsequently, a softmax function is applied to the importance weights. Finally, I viewport candidates can be obtained by

$$\mathbf{v}_i = \mathbf{v}_i^{\text{a}} + \Delta\mathbf{v}_i = \left(\phi_i^{\text{a}} + \Delta\phi_i, \theta_i^{\text{a}} + \Delta\theta_i\right), i \in [1, I]. \quad (1)$$

Viewport softer NMS. Here, We develop a viewport softer NMS for filtering viewport candidates, which is summarized in Algorithm 1. Note that our viewport softer NMS is modified and simplified from the traditional softer NMS on 2D regions [13]. Specifically, among a set of viewport candidates $\mathbf{V}$, the viewport with the maximum importance weight is found. Then, this viewport as well as all viewport candidates within a threshold great-circle distance [16] of d^{th}, are weightily merged according to their importance weights. Similarly, the corresponding importance weights are also merged. These viewport candidates are then excluded from $\mathbf{V}$. Finally, after our viewport softer NMS, the

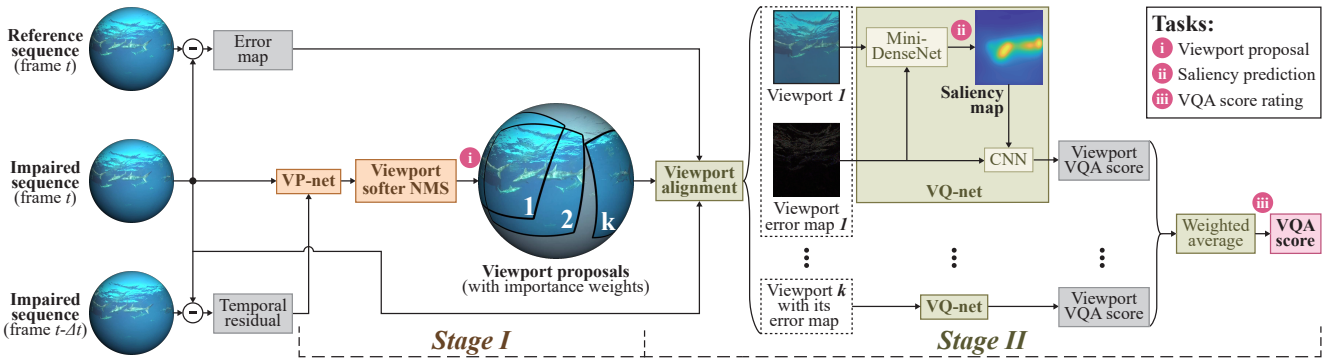


Figure 2. Framework of the V-CNN.

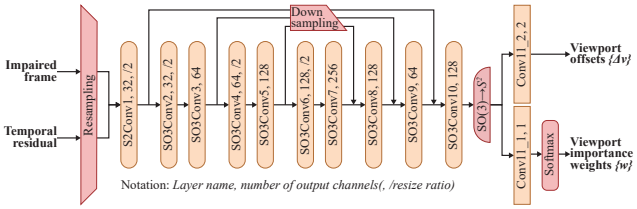


Figure 3. Architecture of the VP-net.

Algorithm 1: Viewport softer NMS.

Input: Sets of viewport candidates $\mathbf{V} = \{\mathbf{v}_1, \ldots, \mathbf{v}_I\}$ and the corresponding predicted importance weights $\mathbf{W} = \{\hat{w}_1, \ldots, \hat{w}_I\}$ at all anchor locations, great-circle distance threshold d^{th}, proposal quantity threshold K^{th}.

Output: Sets of viewport proposals $\mathbf{V}^{\mathrm{P}}$ and the corresponding predicted importance weights $\mathbf{W}^{\mathrm{P}}$.

$k \leftarrow 1$, $\mathbf{V}^{\mathrm{P}} \leftarrow \varnothing$, $\mathbf{W}^{\mathrm{P}} \leftarrow \varnothing$

while $\mathbf{V} \neq \varnothing$ **and** $k \leq K^{\mathrm{th}}$ **do**

  $\iota \leftarrow \arg\max_{\{\iota | \hat{w}_\iota \in \mathbf{W}\}} \hat{w}_\iota$

  $\mathbf{I}' \leftarrow \{\iota' | d(\mathbf{v}_{\iota'}, \mathbf{v}_\iota) < d^{\mathrm{th}}, \mathbf{v}_{\iota'} \in \mathbf{V}\}$, where $d(\mathbf{v}_{\iota'}, \mathbf{v}_\iota)$ is the great-circle distance between $\mathbf{v}_{\iota'}$ and $\mathbf{v}_\iota$

  $w_k^{\mathrm{P}} \leftarrow \sum_{\iota' \in \mathbf{I}'} \hat{w}_{\iota'}$

  $\mathbf{v}_k^{\mathrm{P}} \leftarrow \left(\sum_{\iota' \in \mathbf{I}'} \hat{w}_{\iota'} \cdot \mathbf{v}_{\iota'}\right) / w_k^{\mathrm{P}}$

  $\mathbf{V}^{\mathrm{P}} \leftarrow \mathbf{V}^{\mathrm{P}} \bigcup \{\mathbf{v}_k^{\mathrm{P}}\}$, $\mathbf{W}^{\mathrm{P}} \leftarrow \mathbf{W}^{\mathrm{P}} \bigcup \{w_k^{\mathrm{P}}\}$

  $\mathbf{V} \leftarrow \mathbf{V} \setminus \{\mathbf{v}_{\iota'} | \iota' \in \mathbf{I}'\}$, $\mathbf{W} \leftarrow \mathbf{W} \setminus \{\hat{w}_{\iota'} | \iota' \in \mathbf{I}'\}$

  $k \leftarrow k + 1$

end

return $\mathbf{V}^{\mathrm{P}}, \mathbf{W}^{\mathrm{P}}$

filtered viewports are proposed for the next stage.

3.3. Stage II: VQA

In this stage, the proposed viewports are first projected to 2D plane by our viewport alignment method. Then, the VQA score of each aligned viewport is rated by the proposed VQ-net. Finally, the VQA scores of the proposed viewports at all frames are combined to output the VQA score of an impaired 360° video sequence. More details are presented as follows.

Viewport alignment. Assume that the locations of proposed viewports are $\{\mathbf{v}_k^{\mathrm{P}} = (\phi_k^{\mathrm{P}}, \theta_k^{\mathrm{P}})\}_{k=1}^{K}$, where K is the total number of proposed viewports. At the beginning of stage II, viewport alignment is developed to project the spherical content of the proposed viewport at $\mathbf{v}_k^{\mathrm{P}}$ to the 2D plane, denoted as $\mathbf{C}_k$. For each pixel in $\mathbf{C}_k$, the alignment is implemented by bilinear interpolation on the impaired sequence frame $\mathbf{F}_t$ at the corresponding location. Additionally, the 2D error map of each viewport $\mathbf{C}_k^{\mathrm{err}}$ is also obtained by the same interpolation on the frame error map $\mathbf{F}_t^{\mathrm{err}}$ ($\mathbf{F}_t^{\mathrm{ref}} - \mathbf{F}_t$).

To be more specific, in the 2D plane, we denote the width and height of the viewport as W and H (numbers of pixels), respectively. Given a pixel location (x, y) ($x \in [1, W]$, $y \in [1, H]$) in $\mathbf{C}_k$, the inverse of gnomonic projection [27] is conducted to find the corresponding spherical location $(\phi_{x,y}, \theta_{x,y})$. Firstly, the pixel location (x, y) is converted to (f_x, f_y), which has the same scale as the unit sphere:

$$f_x = \frac{2x-1-W}{W} \cdot \tan\frac{a_W}{2}, \; f_y = -\frac{2y-1-H}{H} \cdot \tan\frac{a_H}{2}, \quad (2)$$

where a_W and a_H are the angular ranges of the viewport, corresponding to W and H, respectively. Then, the corresponding spherical location $(\phi_{x,y}, \theta_{x,y})$ is obtained by

$$\phi_{x,y} = \phi_k^{\mathrm{P}} + \arctan\left(\frac{f_x \sin c}{\rho \cos\theta_k^{\mathrm{P}} \cos c - f_y \sin\theta_k^{\mathrm{P}} \sin c}\right), \quad (3)$$

$$\theta_{x,y} = \arcsin\left(\cos c \sin\theta_k^{\mathrm{P}} + \frac{f_y \sin c \cos\theta_k^{\mathrm{P}}}{\rho}\right), \quad (4)$$

where

$$\rho = \sqrt{f_x^2 + f_y^2}, \quad c = \arctan\rho. \quad (5)$$

Next, the corresponding pixel locations of $(\phi_{x,y}, \theta_{x,y})$ on the $\mathbf{F}_t$ and $\mathbf{F}_t^{\mathrm{err}}$ are determined by the projection type of the frame. Taking the equirectangular projection (ERP) as an example, the corresponding pixel location $(p_{x,y}, q_{x,y})$ of $(\phi_{x,y}, \theta_{x,y})$ can be obtained by:

$$p_{x,y} = \left(\frac{\phi_{x,y}}{360°} + \frac{1}{2}\right) W_{\mathbf{F}} + \frac{1}{2}, \; q_{x,y} = \left(\frac{1}{2} - \frac{\theta_{x,y}}{180°}\right) H_{\mathbf{F}} + \frac{1}{2}, \quad (6)$$

where $W_{\mathbf{F}}$ and $H_{\mathbf{F}}$ are the width and height of the 360° sequence frame, respectively. Finally, the pixel values at

(x, y) in $\mathbf{C}_k$ and $\mathbf{C}_k^{\text{err}}$ are obtained by bilinear interpolation at $(p_{x,y}, q_{x,y})$ in $\mathbf{F}_t$ and $\mathbf{F}_t^{\text{err}}$, yielding the 2D content of proposed viewports.

VQ-net. The architecture of the VQ-net is shown in Figure 4. Both the content $\mathbf{C}_k$ and error map $\mathbf{C}_k^{\text{err}}$ of each proposed viewport are fed into the VQ-net. These 2 inputs are concatenated after going through 2 convolutional layers, respectively. Then, a mini-DenseNet with 3 dense blocks, 3 transpose convolutional layers and a softmax activation are developed to predict the saliency map $\hat{\mathbf{M}}_k$ of each proposed viewport. These saliency maps can be seen as the results of the second auxiliary task for modelling EM of subjects in viewports of 360° video. Next, $\hat{\mathbf{M}}_k$ is up-sampled and then used to weight the input error map $\mathbf{C}_k^{\text{err}}$ by Hadamard product. Subsequently, the weighted error map is fed into a shallow CNN structure with 3 convolutional layers, 2 max pooling layers, a global average pooling layer and 2 fully connected layers. Consequently, the VQA score s_k of each proposed viewport can be obtained.

Final VQA score of 360° video. Given the predicted VQA scores $\{s_k\}_{k=1}^K$ of the proposed viewports, the overall VQA score of one impaired frame (denoted as $\tilde{s}$) can be obtained as the weighted average, according to the importance weights $\{w_k^{\text{p}}\}_{k=1}^K$ of the proposed viewports:

$$\tilde{s} = \frac{\sum_{k=1}^K w_k^{\text{p}} \cdot s_k}{\sum_{k=1}^K w_k^{\text{p}}}. \tag{7}$$

At last, the VQA score of the input impaired 360° video sequence can be obtained by averaging VQA scores of all frames. This is the final output of our V-CNN, as an FR VQA approach on 360° video.

3.4. Training implementation

Now, we focus on the training implementation of our V-CNN approach, in which the VP-net and VQ-net are two CNN models to be trained.

Training VP-net. When watching 360° video, the HM positions of subjects are also the centers of the corresponding viewports. Hence, we use the HM of subjects as the supervision of the predicted importance weights $\hat{w}_i$ and viewport offsets $\Delta\mathbf{v}_i$ in the training of VP-net. Assume that $\{\mathbf{v}_j^{\text{h}} = (\phi_j^{\text{h}}, \theta_j^{\text{h}})\}_{j=1}^J$ are the ground truth HM positions of J subjects. Then, for the i-th anchor location $\mathbf{v}_i^{\text{a}}$, the importance weight w_i can be defined by the following Gaussian distribution and normalization:

$$\tilde{w}_i = \sum_{j=1}^J \exp\left\{-\frac{d^2(\mathbf{v}_i^{\text{a}}, \mathbf{v}_j^{\text{h}})}{2\sigma^2}\right\}, \quad w_i = \frac{\tilde{w}_i}{\sum_{i=1}^I \tilde{w}_i}. \tag{8}$$

In (8), σ is the standard deviation; I is the quantity of anchor locations; $d(\mathbf{v}_i^{\text{a}}, \mathbf{v}_j^{\text{h}})$ is the great-circle distance [16] between $\mathbf{v}_i^{\text{a}}$ and $\mathbf{v}_j^{\text{h}}$. According to (8), importance weights over all anchor locations can be regraded as the probability distribution that whether the corresponding viewport can be watched by subjects. Hence, we use the Kullback-Leibler (KL) divergence as the loss function for predicting the importance weights. To be more specific, we measure the KL divergence D_{KL} between the distribution of importance weights $\{w_i\}_{i=1}^I$ and their predicted values $\{\hat{w}_i\}_{i=1}^I$ at all I anchor locations as follows,

$$\mathcal{L}_w = D_{\text{KL}}(\{w_i\}_{i=1}^I \| \{\hat{w}_i\}_{i=1}^I) = \sum_{i=1}^I w_i \log\left(\frac{w_i}{\hat{w}_i}\right). \tag{9}$$

At each anchor location $\mathbf{v}_i^{\text{a}}$, the ground truth viewport offset $\Delta\mathbf{v}_i^{\text{g}}$ is defined as the difference between $\mathbf{v}_i^{\text{a}}$ and its closest HM position:

$$\Delta\mathbf{v}_i^{\text{g}} = \arg\min_{\{\mathbf{v}_j^{\text{h}}\}_{j=1}^J} d(\mathbf{v}_i^{\text{a}}, \mathbf{v}_j^{\text{h}}) - \mathbf{v}_i^{\text{a}}. \tag{10}$$

Then, a smooth L_1 loss [9], denoted as L_1^{smooth}, is applied for the predicted viewport offset at each anchor location:

$$\mathcal{L}_{\mathbf{v}} = \sum_{i=1}^I w_i \cdot L_1^{\text{smooth}}(\Delta\mathbf{v}_i, \Delta\mathbf{v}_i^{\text{g}}). \tag{11}$$

Note that in (11), the importance weights $\{w_i\}_{i=1}^I$ are utilized to weight L_1^{smooth} of different anchor locations, such that the loss $\mathcal{L}_v$ is more sensitive to the loss at more important anchor locations. Finally, the loss function for training the VP-net is a combination of $\mathcal{L}_w$ and $\mathcal{L}_{\mathbf{v}}$:

$$\mathcal{L}^{\text{I}} = \lambda_w \mathcal{L}_w + \lambda_{\mathbf{v}} \mathcal{L}_{\mathbf{v}}. \tag{12}$$

Training VQ-net. For training the VQ-net, both the outputs of the saliency map and VQA score need to be

Table 1. Values of some key hyper-parameters.

Stage I	Great-circle distance threshold	$d^{\text{th}} = 13.75°$
	Proposal quantity threshold	$K^{\text{th}} = 20$
	Standard deviation in (8)	$\sigma = 18.33°$
	Initial learning rate	1×10^{-3}
	Weight decay	5×10^{-5}
	Coefficient for $\mathcal{L}_w$ in (12)	$\lambda_w = 1$
	Coefficient for $\mathcal{L}_{\mathbf{v}}$ in (12)	$\lambda_{\mathbf{v}} = 5$
Stage II	Viewport size (number of pixels)	$W, H = 540, 600$
	Angular range	$a_W, a_H = 71°, 74°$
	Initial learning rate	1×10^{-3}
	Weight decay	5×10^{-5}
	Coefficient for $\mathcal{L}_{\mathbf{M}_k}$ in (15)	$\lambda_{\mathbf{M}} = 10$
	Coefficient for $\mathcal{L}_{s_k}$ in (15)	$\lambda_s = 1 \times 10^3$

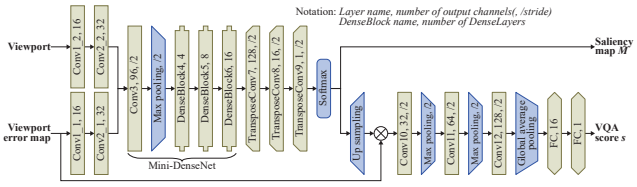


Figure 4. Architecture of the VQ-net.

Table 2. Comparison on VQA performance between our and other approaches, over all test sequences.

Approaches	Attributes				Evaluation on VQA-ODV [19]				
	For 360° video	Full reference	Deep learning	Re-trained*	PLCC	SROCC	KROCC	RMSE	MAE
S-PSNR	✓	✓			0.6929	0.6976	0.4981	8.5407	6.6810
WS-PSNR	✓	✓			0.6721	0.6839	0.4860	8.7707	6.9089
CPP-PSNR	✓	✓			0.6812	0.6896	0.4912	8.6718	6.7932
BP-QAVR	✓	✓	✓	✓	0.6588	0.6801	0.4780	8.9112	7.0823
Li et al.	✓	✓	✓	✓	0.7821	0.7953	0.5902	7.3817	5.7793
VR-IQA-NET	✓		✓		0.3713	0.3379	0.2260	10.9984	9.1016
DeepQA		✓	✓	✓	0.6936	0.7296	0.5213	8.5325	6.7720
WaDIQaM-FR		✓	✓		0.6207	0.6162	0.4206	9.2868	7.4574
V-CNN (Ours)	✓	✓	✓	✓	**0.8740**	**0.8962**	**0.7137**	**5.7551**	**4.4893**

* The training code of the deep learning approaches checked in this column is available. Thus, they are re-trained on the same training set as our approach before evaluation. Other approaches are evaluated with the pre-trained model provided by the authors.

supervised for each input viewport. For the supervision of saliency prediction, we first follow [15] to generate the ground truth saliency map $\mathbf{M}_k$ using the EM points of subjects within the k-th input viewport. Then, the loss of saliency prediction is calculated for the k-th input viewport by the following KL divergence,

$$\mathcal{L}_{\mathbf{M}_k} = D_{\mathrm{KL}}(\mathbf{M}_k \| \hat{\mathbf{M}}_k) = \sum_{(x',y')} M_k(x',y') \log \left\{ \frac{M_k(x',y')}{\hat{M}_k(x',y')} \right\}, \tag{13}$$

where $M_k(x',y')$ and $\hat{M}_k(x',y')$ refer to the saliency values at location (x',y') of $\mathbf{M}_k$ and $\hat{\mathbf{M}}_k$, respectively. For the supervision of VQA, the following loss is applied for the k-th input viewport:

$$\mathcal{L}_{s_k} = (s - s_k)^2, \tag{14}$$

where s_k is the VQA score rated by the VQ-net. In addition, s is the ground truth of difference mean opinion score (DMOS) rated by subjects. Note that it is impractical to obtain the DMOS value of each input viewport. Therefore, in (14), the DMOS value of the 360° sequence is used as the ground truth for supervising the VQA scores of viewports. Finally, the loss function for training the VQ-net is a combination of $\mathcal{L}_{\mathbf{M}_k}$ and $\mathcal{L}_{s_k}$ for the k-th input viewport:

$$\mathcal{L}_k^{\mathrm{II}} = \lambda_{\mathbf{M}} \mathcal{L}_{\mathbf{M}_k} + \lambda_s \mathcal{L}_{s_k}. \tag{15}$$

Training protocol. In training both VP-net and VQ-net, the parameters are updated using the stochastic gradient descent algorithm with the Adam optimizer. Weight decay is also applied for regularization. Table 1 lists the values of some key hyper-parameters in our approach. Among them, the proposal quantity threshold K^{th} is set to 20, according to the recommended number of subjects in subjective VQA on 360° video [33]. Since the typical resolution of viewport in HMD is 1080×1200 with $110 \times 113°$ angular range [8], the standard deviation σ in (8) is set to $18.33°$, such that the range of 3σ can correspond to the angular range of viewport. However, in stage II, the width W and height H of the input viewports are set as half of the resolution of HMD, corresponding to an angular range around $71 \times 74°$. It is because the subjects hardly fix their eyes beyond this range [23]. Moreover, other hyper-parameters are obtained by tuning over the training set.

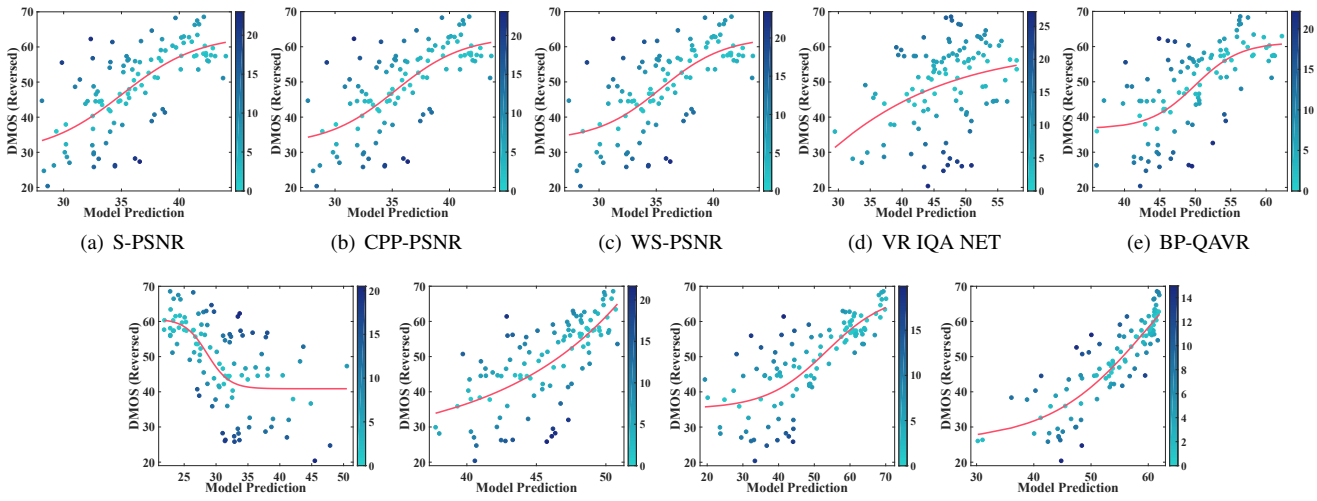
4. Experimental results

4.1. Experiment settings

In this section, we present the experimental results to validate the performance of the proposed V-CNN approach in the main task of VQA on 360° video and in the auxiliary tasks of viewport proposal and viewport saliency prediction. In our experiments, we evaluate the performance in the dataset of VQA-ODV proposed in [19]. Note that the VQA-ODV [19] is the largest VQA dataset for 360° video, which includes the DMOS, HM and EM data of 200+ subjects in watching 540 impaired sequences and 60 reference sequences. We follow the standard training/test set split provided in [19]. That is, our V-CNN model is trained on 432 impaired sequences, and the remaining 108 impaired sequences form the test set. It is worth mentioning that the DMOS values of the VQA-ODV [19] are all reversed in this paper, so that the larger values of DMOS indicate better visual quality.

Then, the performance of VQA can be evaluated by measuring the agreement between the subjective and objective quality scores[2] on the test set. To this end, the correlation between the DMOS values and objective scores of each VQA approach is measured for all test sequences, in terms of 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). We follow [25] to apply a logistic function for fitting the objective VQA scores to their corresponding DMOS values, such that the fitted scores of all VQA approaches are in the same scale as DMOS. Then, PLCC, SROCC, KROCC, RMSE and MAE are calculated for all pairs of the DMOS values and the fitted objective scores. Note that the larger values

[2]Here, objective scores are the predicted results of VQA approaches.



(a) S-PSNR (b) CPP-PSNR (c) WS-PSNR (d) VR IQA NET (e) BP-QAVR

(f) WaDIQaM-FR (g) DeepQA (h) Li *et al.* (i) Our

Figure 5. Scatter plots for the pairs of DMOS values and objective VQA scores of different approaches, over all test sequences.

of PLCC, SROCC and KROCC indicate higher correlation, while the smaller values of RMSE and MAE mean less error of using the objective scores to predict their corresponding DMOS values.

4.2. Evaluation on VQA

Here, we compare the performance between our V-CNN approach and other state-of-the-art approaches for VQA on 360° video, including S-PSNR [38], WS-PSNR [28], CPP-PSNR [39], BP-QAVR [36], VR-IQA-NET [21] and Li [19]. Among them, S-PSNR, WS-PSNR and CPP-PSNR are PSNR-related VQA approaches for 360° video, approved in the standard [37] of the moving picture experts group (MPEG); BP-QAVR, VR-IQA-NET and Li are DNN-based VQA approaches for 360° video. Additionally, we compare two latest DNN-based FR approaches for 2D image: DeepQA [17] and WaDIQaM-FR [2]. For more details about the attributes of these approaches, see Table 2.

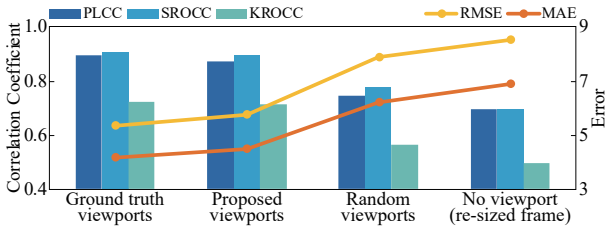
Scatter plots. Figure 5 shows the scatter plots of the objective VQA scores versus the DMOS values over all 108 impaired sequences in the test set. The logistic fitting curves are also shown in this figure. In addition, the color of the scatter points represents the absolute errors to the fitting curves. Generally speaking, the intensive scatter points close to the fitting curves are of little error, indicating the high correlation between the objective VQA scores and their corresponding DMOS values. It can be obviously seen from Figure 5 that the VQA scores of our V-CNN approach have much higher correlation with the DMOS values, compared with all other VQA approaches. Therefore, we can conclude that our V-CNN approach performs much better than other approaches.

Quantitative results. Table 2 tabulates the results of PLCC, SROCC, KROCC, RMSE and MAE for our V-CNN approach and 8 other state-of-the-art approaches. The results of this table are obtained over all test sequences. We can see from Table 2 that our V-CNN approach performs significantly better than all other approaches in terms of 5 metrics. To be more specific, our approach achieves at least 0.09, 0.10, 0.12, 1.62 and 1.28 improvements in PLCC, SROCC, KROCC, RMSE and MAE, respectively.

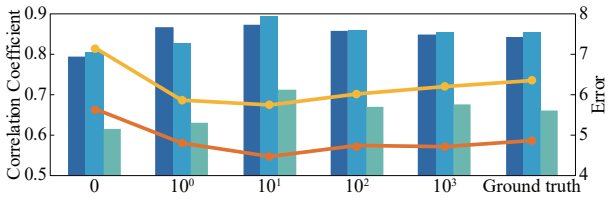
4.3. Ablation study

Ablation on viewport proposal. We investigate the influence of viewport proposal in our V-CNN approach through the following ablation experiments. The experiments include four types of viewports: (1) the ground truth viewports, (2) viewports proposed by our VP-net, (3) viewports at random locations and (4) re-sized frame (regarded as whole video frame but no viewport). Figure 6(a) shows the ablation results on viewports. It can be seen from Figure 6(a) that the performance of VQA is significantly degraded, when the V-CNN approach uses the random viewports or does not incorporate any viewport. In addition, the performance of our V-CNN approach almost remains unchanged, when replacing proposed viewports with the ground-truth viewports. This indicates the effectiveness of the viewport proposal for VQA on 360° video.

Ablation on saliency prediction. We further conduct the ablation experiments to analyze the impact of viewport saliency prediction on VQA. In our experiments, we train the V-CNN with different loss weights on saliency prediction, i.e., $\lambda_{\mathrm{M}} = 0, 10^0, 10^1, 10^2, 10^3$ in (15). Note that larger λ_{M} means more influence of saliency prediction on VQA task. Figure 6(b) shows the VQA performance at different values of λ_{M}. We can see that the V-CNN without saliency prediction (i.e., $\lambda_{\mathrm{M}} = 0$) performs worst. Hence, we can conclude that saliency prediction is effec-



(a) Ablation results on viewport proposal.



(b) Ablation results on saliency prediction.

Figure 6. Results of the ablation experiments.

tive for VQA in our V-CNN approach. On the other hand, the VQA performance slightly reduces, when the V-CNN is trained with smaller or larger λ_{M} than 10^1. This implies that the under- or over-weighted loss on saliency prediction degrades the performance of VQA in our V-CNN approach. In addition, Figure 6(b) also plots the VQA performance when replacing the predicted saliency by its ground truth. As can be seen in this figure, the ground truth saliency leads to a slight reduction on VQA performance, in comparison with that with predicted saliency. It is probably because the Mini-DenseNet of VQ-net learns the shared features for both saliency prediction and VQA, as a typical characteristic of multi-task learning. In summary, the ablation experiments show the saliency prediction has positive impact on our V-CNN approach for VQA.

4.4. Evaluation on auxiliary tasks

In our V-CNN approach, the auxiliary tasks of viewport proposal and viewport saliency prediction can be used to to model the HM and EM of subjects on 360° video, respectively. Here, we verify the performance of our V-CNN approach in these two auxiliary tasks.

Evaluation on viewport proposal. The center of each proposed viewport can be seen as an HM position. We thus evaluate the performance of viewport proposal by generating the heat maps of HM, the same as that in [34]. Then, the accuracy of predicting HM heat maps is measured in three metrics: normalized scanpath saliency (NSS), correlation coefficient (CC) and KL divergence. The larger values of NSS/CC and smaller value of KL divergence mean higher accuracy of HM prediction. For comparison, we evaluate the performance of 4 state-of-the-art HM prediction approaches for 360° video, i.e., Cheng [3], DHP [34], GBVS360 [18] and BMS360 [18]. The experimental results

Table 3. Performance comparison on the auxiliary tasks.

(a) Results of HM modelling

Approaches	NSS	CC	KL
Cheng *et al.*	1.96	0.35	2.50
DHP	1.98	0.37	2.41
GBVS360	1.16	0.22	2.31
BMS360	1.85	0.34	**1.88**
V-CNN (Ours)	**2.65**	**0.63**	2.38

(b) Results of EM modelling

Approaches	NSS	CC	KL
DeepVS	**1.10**	0.36	1.95
BMS	0.89	0.26	1.42
GBVS	0.81	0.30	1.34
PQFT	0.73	0.19	1.60
V-CNN (Ours)	0.97	**0.37**	**1.33**

are shown in Table 3(a). It can be seen from this table that the NSS and CC results of our V-CNN approach are higher than those of all compared approaches, verifying the effectiveness of our V-CNN approach in the first auxiliary task of viewport proposal.

Evaluation on viewport saliency prediction. We further validate the performance of our V-CNN approach in the auxiliary task of viewport saliency prediction, via comparing with 4 state-of-the-art saliency prediction approaches: BMS [40], GBVS [11], PQFT [10] and DeepVS [15]. Similarly, the metrics of NSS, CC and KL are evaluated. The experimental results are reported in Table 3(b). As is shown in this table, our V-CNN approach achieves the best performance in the metrics of CC and KL. Therefore, we can conclude that our V-CNN approach is also effective in the second auxiliary task of viewport saliency prediction.

5. Conclusion

In this paper, we have proposed a V-CNN approach that advances VQA on 360° video by exploring the potential viewports and predicting viewport saliency. Our approach is in accordance with the human's way of watching 360° video, since the observers can only see viewports. Our V-CNN approach consists of two stages, i.e., viewport proposal and VQA. In the first stage, the VP-net was developed with a viewport softer NMS for proposing potential viewports. In this stage, the first auxiliary task of viewport proposal is achieved. In the second stage, we designed a VQ-net to rate the VQA score of each proposed viewport. In the VQ-net, the saliency map of each proposed viewport is predicted, as the second auxiliary task; then it is used for VQA of the viewport. Subsequently, the main task of VQA on 360° video is accomplished by integrating the VQA scores of all proposed viewports. Finally, the experimental results verified that the V-CNN approach significantly outperforms 8 other state-of-the-art approaches in VQA on 360° video, and that it also performs well in the auxiliary tasks of viewport proposal and viewport saliency prediction.

Acknowledgement

This work was supported by NSFC under Grants 61876013 and 61573037, and by the Fok Ying-Tung Education Foundation under Grant 151061.

References

[1] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Saltinet: Scan-path prediction on 360 degree images using saliency volumes. In *IEEE International Conference on Computer Vision Workshop*, pages 2331–2338. IEEE, 2017.

[2] Sebastian Bosse, Dominique Maniry, Klaus-Robert Müller, Thomas Wiegand, and Wojciech Samek. Deep neural networks for no-reference and full-reference image quality assessment. *IEEE Transactions on Image Processing*, 27(1):206–219, 2018.

[3] Hsien-Tzu Cheng, Chun-Hung Chao, Jin-Dong Dong, Hao-Kai Wen, Tyng-Luh Liu, and Min Sun. Cube padding for weakly-supervised saliency prediction in 360° videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1420–1429. IEEE, 2018.

[4] B Choi, YK Wang, MM Hannuksela, Y Lim, and A Murtaza. Information technology–coded representation of immersive media (MPEG-I)–part 2: Omnidirectional media format. *ISO/IEC*, pages 23090–2, 2017.

[5] Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *International Conference on Learning Representations*, 2018.

[6] Xavier Corbillon, Francesca De Simone, and Gwendal Simon. 360-degree video head movement dataset. In *ACM International Conference on Multimedia Systems*, pages 199–204. ACM, 2017.

[7] Erwan J David, Jesús Gutiérrez, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. A dataset of head and eye movements for 360° videos. In *ACM International Conference on Multimedia Systems*, pages 432–437. ACM, 2018.

[8] Alex Davies. Oculus Rift vs. HTC Vive vs. PlayStation VR. `https://www.tomshardware.co.uk/vive-rift-playstation-vr-comparison,review-33556-3.html/`, 2016.

[9] Ross Girshick. Fast R-CNN. In *IEEE International Conference on Computer Vision*, pages 1440–1448, 2015.

[10] Chenlei Guo, Qi Ma, and Liming Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *IEEE Conference on Computer Vision and Pattern Recognition*. Citeseer, 2008.

[11] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In *Advances in Neural Information Processing Systems*, pages 545–552, 2007.

[12] Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross Girshick. Mask R-CNN. In *IEEE International Conference on Computer Vision*, pages 2980–2988. IEEE, 2017.

[13] Yihui He, Xiangyu Zhang, Marios Savvides, and Kris Kitani. Softer-NMS: Rethinking bounding box regression for accurate object detection. *arXiv preprint arXiv:1809.08545*, 2018.

[14] Dennis M Healy, Daniel N Rockmore, Peter J Kostelec, and Sean Moore. FFTs for the 2-sphere-improvements and variations. *Journal of Fourier Analysis and Applications*, 9(4):341–385, 2003.

[15] Lai Jiang, Mai Xu, Tie Liu, Minglang Qiao, and Zulin Wang. DeepVS: A deep learning based video saliency prediction approach. In *European Conference on Computer Vision*, pages 625–642. Springer, 2018.

[16] Lyman M Kells. *Plane and Spherical Trigonometry with Tables by Lyman M. Kells, Willis F. Kern, James R. Bland*. US Armed Forces Institute, 1940.

[17] Jongyoo Kim and Sanghoon Lee. Deep learning of human visual sensitivity in image quality assessment framework. In *IEEE Conference on Computer Vision and Pattern Recognition*, 2017.

[18] Pierre Lebreton and Alexander Raake. GBVS360, BMS360, ProSal: Extending existing saliency prediction models from 2D to omnidirectional images. *Signal Processing: Image Communication*, 2018.

[19] Chen Li, Mai Xu, Xinzhe Du, and Zulin Wang. Bridge the gap between VQA and human behavior on omnidirectional video: A large-scale dataset and a deep learning model. In *ACM Multimedia Conference*, pages 932–940. ACM, 2018.

[20] Feng Li, Huihui Bai, and Yao Zhao. Visual attention guided eye movements for 360 degree images. In *IEEE Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*, pages 506–511. IEEE, 2017.

[21] Heountaek Lim, Hak Gu Kim, and Yang Man Ra. VR IQA NET: Deep virtual reality image quality assessment using adversarial learning. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 6737–6741, April 2018.

[22] Francisco Lopes, João Ascenso, António Rodrigues, and Maria Paula Queluz. Subjective and objective quality assessment of omnidirectional video. In *Applications of Digital Image Processing XLI*, volume 10752, page 107520P. International Society for Optics and Photonics, 2018.

[23] Yashas Rai, Patrick Le Callet, and Philippe Guillotel. Which saliency weighting for omni directional image quality assessment? In *IEEE International Conference on Quality of Multimedia Experience*, pages 1–6. IEEE, 2017.

[24] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster R-CNN: towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (6):1137–1149, 2017.

[25] Kalpana Seshadrinathan, Rajiv Soundararajan, Alan Conrad Bovik, and Lawrence K Cormack. Study of subjective and objective quality assessment of video. *IEEE Transactions on Image Processing*, 19(6):1427–1441, 2010.

[26] 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*, 24(4):1633–1642, 2018.

[27] John Parr Snyder. *Map projections–A working manual*, volume 1395. US Government Printing Office, 1987.

[28] Yule Sun, Ang Lu, and Lu Yu. Weighted-to-spherically-uniform quality evaluation for omnidirectional video. *IEEE Signal Processing Letters*, 24(9):1408–1412, 2017.

[29] Huyen TT Tran, Nam Pham Ngoc, Cuong Manh Bui, Minh Hong Pham, and Truong Cong Thang. An evaluation of quality metrics for 360 videos. In *IEEE International*

Conference on Ubiquitous and Future Networks, pages 7–11. IEEE, 2017.

[30] Evgeniy Upenik, Martin Rerabek, and Touradj Ebrahimi. On the performance of objective metrics for omnidirectional visual content. In *IEEE International Conference on Quality of Multimedia Experience*, pages 1–6. IEEE, 2017.

[31] Chenglei Wu, Zhihao Tan, Zhi Wang, and Shiqiang Yang. A dataset for exploring user behaviors in VR spherical video streaming. In *ACM International Conference on Multimedia Systems*, pages 193–198. ACM, 2017.

[32] Xiaoyu Xiu, Yuwen He, Yan Ye, and Bharath Vishwanath. An evaluation framework for 360-degree video compression. In *IEEE International Conference on Visual Communications and Image Processing*, pages 1–4. IEEE, 2017.

[33] Mai Xu, Chen Li, Yufan Liu, Xin Deng, and Jiaxin Lu. A subjective visual quality assessment method of panoramic videos. In *IEEE International Conference on Multimedia and Expo*, pages 517–522. IEEE, 2017.

[34] Mai Xu, Yuhang Song, Jianyi Wang, MingLang Qiao, Liangyu Huo, and Zulin Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[35] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao. Gaze prediction in dynamic 360° immersive videos. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 5333–5342, 2018.

[36] Shu Yang, Junzhe Zhao, Tingting Jiang, Jing Wang Tariq Rahim, Bo Zhang, Zhaoji Xu, and Zesong Fei. An objective assessment method based on multi-level factors for panoramic videos. In *IEEE International Conference on Visual Communications and Image Processing*, pages 1–4. IEEE, 2017.

[37] 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*, 16, 2017.

[38] Matt Yu, Haricharan Lakshman, and Bernd Girod. A framework to evaluate omnidirectional video coding schemes. In *IEEE International Symposium on Mixed and Augmented Reality*, pages 31–36. IEEE, 2015.

[39] Vladyslav Zakharchenko, Kwang Pyo Choi, and Jeong Hoon Park. Quality metric for spherical panoramic video. In *Optics and Photonics for Information Processing X*, volume 9970, page 99700C. International Society for Optics and Photonics, 2016.

[40] Jianming Zhang and Stan Sclaroff. Exploiting surroundedness for saliency detection: a boolean map approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, (5):889–902, 2016.