Learned Scanpaths Aid Blind Panoramic Video Quality Assessment

Kanglong Fan[1], Wen Wen[1], Mu Li[2]*, Yifan Peng[3], and Kede Ma[1]
[1] City University of Hong Kong, [2] Harbin Institute of Technology, Shenzhen
[3] The University of Hong Kong
{kanglofan2-c,wwen29-c}@my.cityu.edu.hk, limuhit@gmail.com
evanpeng@hku.hk, kede.ma@cityu.edu.hk
https://github.com/kalofan/AutoScanpathQA

Abstract

Panoramic videos have the advantage of providing an immersive and interactive viewing experience. Nevertheless, their spherical nature gives rise to various and uncertain user viewing behaviors, which poses significant challenges for panoramic video quality assessment (PVQA). In this work, we propose an end-to-end optimized, blind PVQA method with explicit modeling of user viewing patterns through visual scanpaths. Our method consists of two modules: a scanpath generator and a quality assessor. The scanpath generator is initially trained to predict future scanpaths by minimizing their expected code length and then jointly optimized with the quality assessor for quality prediction. Our blind PVQA method enables direct quality assessment of panoramic images by treating them as videos composed of identical frames. Experiments on three public panoramic image and video quality datasets, encompassing both synthetic and authentic distortions, validate the superiority of our blind PVQA model over existing methods.

1. Introduction

The rapid advancement of multimedia technologies has marked the beginning of a new era characterized by the proliferation of panoramic videos [20]. Such type of digital data offers an immersive and interactive viewing experience that is transforming the way we consume multimedia. Therefore, assessing and ensuring the visual quality of panoramic videos is increasingly important, as it shapes the users' viewing experience and the triumph of any product or service based on panoramic videos [40]. Unlike their planar counterparts, panoramic videos provide a 360° broad view with a spherical data structure, which poses significant computational challenges for panoramic video quality assessment (PVQA). Moreover, the diverse and uncertain user viewing behaviors (in the form of visual scanpaths) induced by the spherical structure further complicate the quality prediction process. Addressing these challenges requires novel PVQA models that take into account both the spherical data structure as well as user viewing patterns.

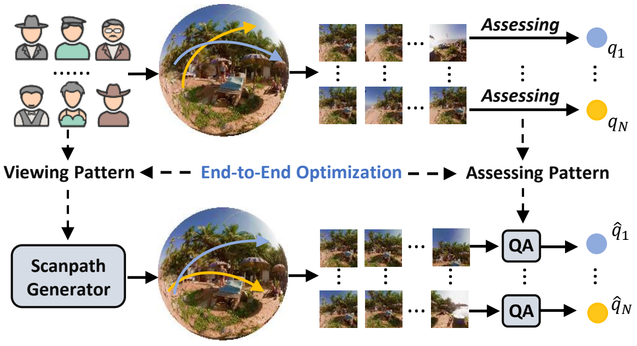

Figure 1. Analogy between human subjects and our end-to-end optimized method for panoramic video quality assessment.

In the quality assessment of panoramic images and videos, three approaches are commonly employed: sphere-to-plane projection onto a 2D plane, rectilinear projection onto multiple viewports, and direct processing using spherical operations. According to the Theorema Egregium by Gauss, all sphere-to-plane map projections [32, 35, 41] are impeded by non-uniform sampling and geometric distortions, which may bias subsequent *planar* quality prediction. While spherical operators give a better account for the panoramic data structure, they are generally computationally prohibitive and, more importantly, may not faithfully reflect user viewing patterns [4, 39, 40]. To overcome these computational difficulties, several methods seek to sample and process rectilinear viewports [7, 14, 27, 29, 36, 37]. Of particular interest are scanpath-based methods, which sample, along visual scanpaths [23, 24], sequences of rectilinear viewports at discrete time instances. This sampling process turns panoramic images and videos into moving-

*Corresponding author.

camera videos, amenable to *planar* VQA.

By closely imitating how humans perceive visual distortions in virtual environments (see Figure 1), scanpath-based methods [27, 29, 36] have demonstrated remarkable efficacy in the quality of panoramic images. Nonetheless, some methods [27] rely on human visual scanpaths for assessment, which are cumbersome and time-consuming to obtain, thus limiting their applications in fully-automated situations. Some other methods [29, 36] design and refine the scanpath generator separately from the quality predictor, which is bound to be suboptimal. Moreover, while all methods prove effective with panoramic images, their adaptability for use with panoramic videos remains unclear.

In this work, we further pursue the scanpath-based methods for end-to-end optimized blind PVQA. Our method consists of two modules: a scanpath generator and a quality assessor. Our scanpath generator is probabilistic, which takes historical scanpaths as input and is pre-trained to predict future scanpaths by minimizing their expected code length [16]. The scanpath generator and the quality assessor are then jointly optimized to explain human perceptual scores of panoramic videos. To enable end-to-end optimization, we employ the reparameterization trick [12, 13] to allow differentiable scanpath sampling and adopt subgradients to handle discontinuities of the interpolation kernel [11] for viewport sequence generation. Our blind PVQA method not only eliminates the need for human scanpaths, but also supplies a lightweight and differentiable scanpath generator that can work with any planar VQA model. Furthermore, our method is “backward compatible,” in the sense that it handles panoramic images with no modification. We test the proposed blind PVQA models on three public panoramic image and video quality datasets [6, 30, 35], covering both synthetic and authentic distortions. Under both in-dataset and cross-dataset settings, our models consistently exhibit better quality prediction performance.

2. Related Work

We review two highly relevant topics, scanpath generation and quality assessment of panoramic images and videos.

2.1. Scanpath Generation

Typical inputs to a panoramic scanpath generator include the saliency map, optical flow map, and historical scanpath. To improve saliency detection, Nguyen *et al.* [21] compiled a panoramic video saliency dataset, while Xu *et al.* [38] focused on relative viewpoint displacement prediction. Apart from the historical scanpath, Li *et al.* [15] incorporated “future” scanpaths from other users to facilitate cross-user transfer learning. Through an in-depth root-cause analysis, Rondón *et al.* [25] discovered that visual features have a minimal impact on the prediction of short-term scanpaths (*e.g.*, $\leq$ 2 seconds). Motivated by their findings, Chao *et al.* [3] trained a Transformer [33] to predict future scanpaths based solely on historical scanpaths.

The above-mentioned methods [3, 15, 21, 25, 38] treat scanpath generation as a deterministic prediction task, neglecting the inherent scanpath diversity and uncertainty. As a departure, Li *et al.* [16] formulated scanpath generation as a density estimation problem, which can be implemented by expected code length minimization. In our work, we adopt Li’s approach [16] to learn multi-user viewing patterns and generate human-like scanpaths.

2.2. Quality Assessment

Current PVQA models are primarily derived from planar image and video quality methods, which are applied to three types of data representations: the projected 2D plane, spherical surface, and projected rectilinear viewport.

Planar domain methods [32, 35, 41] aim to rectify geometric distortions and mitigate uneven sampling that results from the sphere-to-plane projection. These include the latitude-adaptive weighting [32], Craster parabolic projection [41], and pseudocylindrical representation [35]. Spherical domain methods, such as S-PSNR [40] and S-SSIM [4], compute and pool local quality measurements over the sphere. Yang *et al.* [39] trained a non-local spherical neural network [5, 34] to extract spatiotemporal information from panoramic videos. Viewport domain methods prioritize the extraction of visually informative viewports for quality analysis. Li *et al.* [14] introduced a two-step approach that involves viewport proposal and quality assessment. Xu *et al.* [37] built a graph over the extracted viewports, and Fu *et al.* [7] constructed hypergraphs to represent the semantic interactions between viewports. One limitation of current viewport proposal methods is that they do not accurately reflect the human viewing experience.

Sui *et al.* [27] pioneered scanpath-based methods for PVQA, under the category of viewport domain methods. To eliminate the dependency on human scanpaths, Sui *et al.* [29] adopted a deep Markov model [28] to generate scanpaths. Meanwhile, Wu *et al.* [36] handcrafted a simple scanpath generator based on the entropy feature and equator bias. These methods are tailored for panoramic images and are not end-to-end optimized. In contrast, we aim ambitiously for an end-to-end optimized quality assessment method for panoramic videos, with the added benefit of being backward compatible with panoramic images.

3. Proposed Method

As illustrated in Figure 2, our method consists of two modules: a scanpath generator and a quality assessor. Given a panoramic video, we first specify a starting point (ϕ_0, θ_0), a viewing duration S, and N initial paths. The scanpath generator autoregressively samples N scanpaths based on the

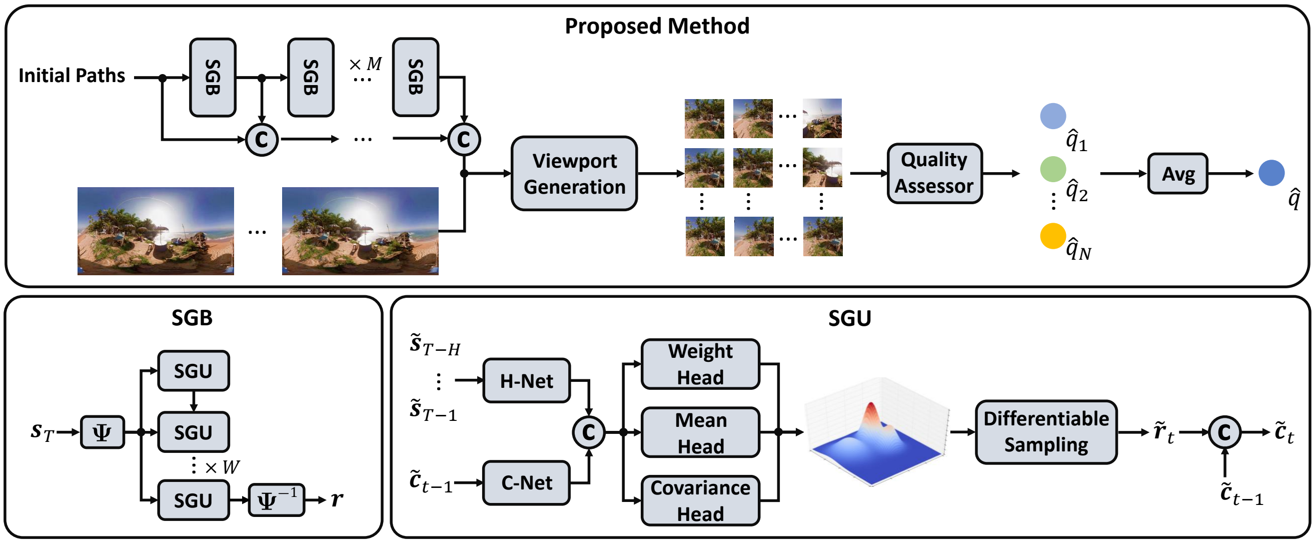

Figure 2. Overview of the proposed blind PVQA method, consisting of a scanpath generator and a quality assessor. The basic component of the scanpath generator is the scanpath generation unit (SGU), which utilizes the historical and causal relative scanpaths to produce the GMM parameters for differentiable sampling of the current viewpoint. By assembling W SGUs, we create a scanpath generation block (SGB), which autoregressively predicts a future scanpath of W viewpoints. We further stack M SGBs to generate a long-term scanpath of $M \times W + H$ viewpoints, where H is the length of the initial path. By adjusting the number of initial paths (denoted by N), we can sample N scanpaths, along which we produce N viewport sequences as input to the quality assessor.

initial and already generated path segments. Along these scanpaths, we apply a differentiable viewport generation technique to extract N viewport sequences from the input panoramic video. Each viewport sequence (as a planar video) is fed to the quality assessor, whose predicted score is subsequently aggregated into an overall quality estimate of the panoramic video.

3.1. Scanpath Generator

Probabilistic Scanpath Modeling. To capture the uncertainty and diversity of human scanpaths, we formulate panoramic scanpath generation as a density estimation problem:

$$\max p(\boldsymbol{r}|\boldsymbol{s}), \tag{1}$$

where $\boldsymbol{s} = \{(\phi_0, \theta_0), \ldots, (\phi_t, \theta_t), \ldots, (\phi_{T-1}, \theta_{T-1})\}$ is the historical scanpath as the condition and $\boldsymbol{r} = \{(\phi_T, \theta_T), \ldots, (\phi_{T+W-1}, \theta_{T+W-1})\}$ is the future scanpath to be predicted. Herein, W is the prediction horizon, and (ϕ_t, θ_t) is the t-th viewpoint in the Euler coordinate system. Mathematically, $p(\boldsymbol{r}|\boldsymbol{s})$ can be decomposed as

$$p(\boldsymbol{r}|\boldsymbol{s}) = \prod_{t=0}^{W-1} p\left(\phi_{T+t}, \theta_{T+t}\middle|\boldsymbol{s}, \boldsymbol{c}_t\right), \tag{2}$$

where $\boldsymbol{c}_t = \{(\phi_T, \theta_T), \ldots, (\phi_{T+t-1}, \theta_{T+t-1})\}$ is referred to as the causal path that includes all estimated viewpoints before $(\phi_{T+t}, \theta_{T+t})$, and $\boldsymbol{c}_0 = \emptyset$. The chain rule suggests estimating the conditional probability $p\left(\phi_{T+t}, \theta_{T+t}\middle|\boldsymbol{s}, \boldsymbol{c}_t\right)$ autoregressively. We further make the Markovian assumption: prediction of the current viewpoint is conditionally independent of viewpoints that are temporally further distant, given the most recent H viewpoints. This leads to a truncated historical path context $\boldsymbol{s}_T = \{(\phi_{T-H}, \theta_{T-H}), \ldots, (\phi_{T-1}, \theta_{T-1})\}$.

We parameterize the probability $p\left(\boldsymbol{r}_t\middle|\boldsymbol{s}_T, \boldsymbol{c}_t\right)$, where $\boldsymbol{r}_t = (\phi_{T+t}, \theta_{T+t})$, by a Gaussian mixture model (GMM) with K components:

$$p\left(\boldsymbol{r}_t\middle|\boldsymbol{s}_T, \boldsymbol{c}_t\right) = \sum_{i=1}^{K} \alpha_i \mathcal{N}_i(\boldsymbol{r}_t; \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i), \tag{3}$$

where α_i is the i-th mixture weight, $\boldsymbol{\mu}_i$ and $\boldsymbol{\Sigma}_i$ represent the mean vector and the covariance matrix of the i-th Gaussian component, respectively. This parametrization can be straightforwardly done by training a density estimation network for parameter estimation. As illustrated in Figure 2, this network is inside the scanpath generation unit (SGU) and is composed of two subnetworks to process the historical path context $\boldsymbol{s}_T$ and the causal path context $\boldsymbol{c}_t$, which we denote by H-Net and C-Net, respectively. The concatenated features are fed to three prediction heads to estimate the weight vector, the mean vectors, and the covariance matrices of the GMM, respectively. We find empirically that incorporating the historical video frames as the visual context significantly increases computational demands with only slight improvements in performance. Therefore, to keep the scanpath generator lightweight, we choose to omit the visual context. The detailed specifications of the density estimation network can be found in the supplementary material.

Estimating continuous probability density is generally difficult and may lead to overfitting. In particular, maximum likelihood estimation of the GMM parameters through direct optimization of Eq. (3) is challenging, due to the presence of singularities [2]. To circumvent this, we compute the probability mass $P\left(\bar{\boldsymbol{r}}_t \middle| \boldsymbol{s}_T, \boldsymbol{c}_t\right)$ by discretizing and integrating the density $p\left(\boldsymbol{r}_t \middle| \boldsymbol{s}_T, \boldsymbol{c}_t\right)$:

$$P\left(\bar{\boldsymbol{r}}_t | \boldsymbol{s}_T, \boldsymbol{c}_t\right) = \int_{\Omega} p\left(\bar{\boldsymbol{r}}_t | \boldsymbol{s}_T, \boldsymbol{c}_t\right) d\Omega. \tag{4}$$

$\bar{\boldsymbol{r}}_t$ represents the quantized value of $\boldsymbol{r}_t$ by a uniform quantizer with a step size of Δ:

$$\bar{\xi} = Q(\xi) = \Delta \left\lfloor \frac{\xi}{\Delta} + \frac{1}{2} \right\rfloor, \tag{5}$$

where $\lfloor \cdot \rfloor$ denotes the floor function. $\Omega = [\bar{\phi}_{T+t} - 1/2\Delta, \bar{\phi}_{T+t} + 1/2\Delta] \times [\bar{\theta}_{T+t} - 1/2\Delta, \bar{\theta}_{T+t} + 1/2\Delta]$ is the integration interval. As pointed out in [16], the incorporation of quantization establishes the equivalence between scanpath generation and lossy scanpath compression.

Furthermore, the absolute Euler coordinate system is not user-centric, meaning that it is not centered at the user's current viewpoint, relative to historical and future viewpoints [16]. This may complicate the probabilistic modeling of scanpaths and the end-to-end optimization of blind PVQA. To address this, we convert the Euler coordinates to the relative uv coordinates:

$$\tilde{\boldsymbol{s}}_{T-t} = \boldsymbol{\Psi}_{T-t}(\boldsymbol{s}_T), \text{ for } t \in \{1, \ldots, H\}, \tag{6}$$

where $\boldsymbol{\Psi}_{T-t}(\cdot)$ denotes the mapping of $\boldsymbol{s}_T$ to the viewport centered at the reference viewpoint $(\phi_{T-t}, \theta_{T-t})$. By choosing each viewpoint in $\boldsymbol{s}_T$ as the reference, we create H relative scanpaths out of $\boldsymbol{s}_T$ (see Figure 3), which serve as input to the H-Net. Meanwhile, we map the causal scanpath context $\boldsymbol{c}_t$ and the viewpoint to be predicted $(\phi_{T+t}, \theta_{T+t})$ to the last historical viewport centered at $(\phi_{T-1}, \theta_{T-1})$. We stack W SGUs to form a scanpath generation block (SGB), which takes $\boldsymbol{s}_T$ as input and predicts the future scanpath $\boldsymbol{r}$. Furthermore, we stack M SGBs to form the scanpath generator, which is capable of predicting a very long-term scanpath of length $M \times W + H$ (including the initial length H). The parameters of different SGUs are shared to enable variable-length scanpath generation by varying M.

Differentiable Scanpath Sampling. To enable end-to-end optimization of the proposed blind PVQA method, we propose a two-step differentiable sampling method to draw viewpoints from the estimated GMM via the reparameterization trick [12, 13]. The first step is to select a Gaussian component from which to sample the viewpoint, according to the categorical distribution:

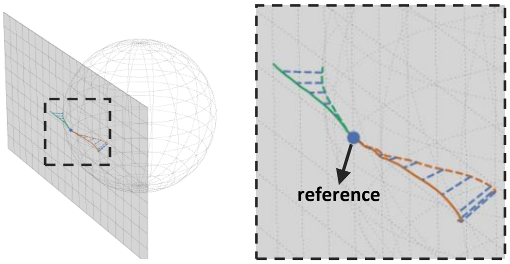

Figure 3. Visualization of a relative scanpath projected from the sphere to the viewport.

$$\boldsymbol{e} = \text{one\_hot}\left(\mathop{\arg\max}_{i \in \{1, \ldots K\}} (\log(\alpha_i) + g_i)\right), \tag{7}$$

where g_i is a sample drawn from the $\text{Gumbel}(0, 1)$ distribution, and $\boldsymbol{e}$ is a one-hot vector. Eq. (7) is known as the Gumbel-Max trick [8], which is non-differentiable. We relax the $\arg\max$ operator with a softmax function [12]:

$$\hat{\boldsymbol{e}} = \text{softmax}((\log(\boldsymbol{\alpha}) + \boldsymbol{g})/\tau), \tag{8}$$

where τ represents the temperature coefficient, $\boldsymbol{\alpha} = [\alpha_1, \ldots, \alpha_K]$, and $\boldsymbol{g} = [g_1, \ldots, g_K]$. As τ approaches zero, $\hat{\boldsymbol{e}}$ converges to $\boldsymbol{e}$. In the forward pass, $\arg\max$ is used directly, while in the backward pass, it is replaced by the softmax function . The second step involves sampling a viewpoint from the selected Gaussian component. Assuming the i-th Gaussian component is selected, the linear reparameterization trick [13] suggests

$$\tilde{\boldsymbol{r}}_t = \boldsymbol{u}_i + \boldsymbol{L}_i \boldsymbol{\epsilon}, \tag{9}$$

where $\tilde{\boldsymbol{r}}_t$ is the relative uv coordinates of the t-th future viewpoint. $\boldsymbol{\Sigma}_i = \boldsymbol{L}_i \boldsymbol{L}_i^\intercal$ is the Cholesky decomposition, and $\boldsymbol{\epsilon}$ is a sample drawn from the $\mathcal{N}(\boldsymbol{0}, \mathbf{I})$. Differentiation of the Cholesky decomposition is complicated and sometimes numerically unstable [26]. Thus we assume the independence between the uv coordinates, leading to the simplified reparameterization formula:

$$\tilde{\boldsymbol{r}}_t = \boldsymbol{u}_i + \boldsymbol{\sigma}_i \odot \boldsymbol{\epsilon}, \tag{10}$$

where $\odot$ denotes the element-wise product, and $\boldsymbol{\sigma}_i$ denotes the standard deviations of the i-th Gaussian component. Through the two-step reparameterization, our sampling strategy ensures effective back-propagation. As suggested in [16], we additionally implement a proportional–integral–derivative (PID) controller [1] to further improve the smoothness of the sampled scanpaths (see the details in the supplementary material).

Differentiable Viewport Sequence Generation. Inspired by [11], we first parameterize the Euler sampling grid in terms of the relative uv coordinates via the inverse transformation $\mathbf{\Psi}^{-1}$ (see Eq. (6)). Subsequently, the Euler coordinate (ϕ, θ) is mapped to the discrete sampling position (m, n) in the ERP domain:

$$m = (0.5 - \phi/\pi)H_e - 0.5, \tag{11}$$

$$n = (\theta/2\pi + 0.5)W_e - 0.5, \tag{12}$$

where H_e and W_e are the height and width of the video frame in ERP. Once the mapping between (u, v) and (m, n) is established, we construct a flow field that is the same size as the viewport. Within this field, each element records the corresponding pixel position (m, n). Given an ERP image and the flow field, we apply bilinear interpolation [11] to compute pixel values in the viewport and leverage its sub-gradients for back-propagation. This process yields N viewport sequences, corresponding to N initial paths.

3.2. Quality Assessor

Our probabilistic scanpath generator can work with any planar VQA model, whether it is differentiable or not. To enable end-to-end optimization of the scanpath generator and the quality assessor, and to make a fair comparison with existing blind PVQA models, we reuse three differentiable quality assessors from ScanpathVQA [35], GSR-S / GSR-X [29], and Assessor360 [36]. Specifically, the quality assessor of ScanpathVQA is a lightweight ResNet-18 network [9], with the classification head replaced by a quality estimator (a multilayer perceptron). The quality assessors of GSR-S / GSR-X are adapted from Video Swin-T [18] / X-Clip-B/32 [22]. The quality assessor of Assessor360 is modified from Swin-B [17] with the addition of a temporal analysis module. We feed each of the N viewport sequences to the quality assessor to compute N quality scores. The overall quality estimate is then computed by a simple average:

$$\hat{q} = \frac{1}{N}\sum_{i=1}^{N}\hat{q}_i. \tag{13}$$

3.3. Optimization Strategy

We explore a three-stage training procedure for our blind PVQA model. In the first stage, we pre-train the density estimation network on the VRVQW dataset [35] by minimizing the expected code length of the generated scanpaths (also equivalent to minimizing the negative log-likelihood):

$$\ell_{\text{code}} = -\frac{1}{BW}\sum_{i=1}^{B}\sum_{t=0}^{W-1}\log_2\left(P\left(\bar{\boldsymbol{r}}_t^{(i)}\middle|\boldsymbol{s}^{(i)}, \boldsymbol{c}_t^{(i)}\right)\right), \tag{14}$$

where B denotes the mini-batch size. During this stage of training, we use human scanpaths to fill in the causal path context, which can be efficiently implemented by a causal masking mechanism. In the second stage, we fix the parameters of the pre-trained scanpath generator and warm up the quality assessor by optimizing the Pearson linear correlation coefficient (PLCC) between human perceptual scores and model predictions. In the third stage, we end-to-end finetune the entire method. We find that the proposed three-stage optimization strategy accelerates convergence compared to the naive end-to-end optimization.

4. Experiments

In this section, we first delineate the experimental setups, and then compare our method with current blind PVQA models under both in-dataset and cross-dataset settings. We further validate our scanpath generator in terms of explaining human perceptual scores and replicating human viewing patterns. Lastly, we conduct a series of ablation experiments to probe the impact of several key designs.

4.1. Experimental Setups

Datasets. We employ three panoramic image and video datasets: VRVQW [35], CVIQD [30], and OIQA [6]. The VRVQW dataset includes 502 *panoramic videos* that have a wide spectrum of authentic distortions. Each video is viewed under four unique viewing conditions to simulate the different quality of experience during the initial viewing. The CVIQD dataset comprises a total of 528 compressed *panoramic images* by JPEG, AVC, and HEVC, from 16 reference images. The OIQA dataset includes 320 *panoramic images* that have been altered from 16 reference images by JPEG compression, JPEG2000 compression, Gaussian blurring, and Gaussian noise contamination.

Implementation Details. For the scanpath generator, we set the length of the provided initial path H and the predicted future path W in the SGB to be identical and equal to 5. The number of Gaussian components K in Eq. (3) is set to 3. The quantization step size Δ in Eq. (5) is set to 0.2. The temperature coefficient τ in Eq. (8) is set to 1. The number of stacked SGBs M is set to 6 and 14 for the viewing duration of 7 and 15 seconds, respectively. For the quality assessor, the number of scanpaths N in Eq. (13) is set to 20. The input viewport size $H_v \times W_v$ is 224×224, corresponding to a field of view of $90^\circ \times 90^\circ$. The length of the viewport sequence is set to 7 regardless of the duration and frame rate of the original panoramic video. We split VRVQW randomly into the training, validation, and test sets according to the ratio of $6 : 2 : 2$ for 5 times, and report the mean results. Similarly, we split CVIQD and OIQA using a different ratio of $7 : 1 : 2$ for 5 times. The detailed configuration of our three-stage optimization strategy is detailed in the supplementary material.

Table 1. In-dataset comparison of blind PVQA methods on three panoramic image and video quality datasets. SRCC: Spearman's rank correlation coefficient. PLCC: Pearson linear correlation coefficient. The same evaluation metrics are applied in Tables 2, 3, 5, 6 and Figure 6. The best results on each dataset are highlighted in bold.

Dataset	Method	SRCC	PLCC
VRVQW [35]	NIQE [19]	0.401	0.365
	MC360IQA [31]	0.669	0.671
	Wen24 [35]	0.756	0.763
	ScanpathVQA [35]	0.779	0.781
	Assessor360 [36]	0.406	0.415
	Ours (ScanpathVQA)	0.801	0.809
	Ours (GSR-S)	0.804	0.807
	Ours (GSR-X)	0.815	0.819
	Ours (Assessor360)	**0.822**	**0.823**
CVIQD [30]	NIQE [19]	0.847	0.878
	MC360IQA [31]	0.917	0.939
	Wen24 [35]	0.919	0.932
	GSR-S [29]	0.905	0.937
	GSR-X [29]	0.944	0.962
	Assessor360 [36]	0.955	0.969
	Ours (ScanpathVQA)	0.912	0.936
	Ours (GSR-S)	0.930	0.958
	Ours (GSR-X)	0.956	0.974
	Ours (Assessor360)	**0.972**	**0.983**
OIQA [6]	NIQE [19]	0.702	0.657
	MC360IQA [31]	0.900	0.906
	Wen24 [35]	0.905	0.907
	GSR-S [29]	0.902	0.915
	GSR-X [29]	0.945	0.954
	Assessor360 [36]	0.946	0.953
	Ours (ScanpathVQA)	0.915	0.922
	Ours (GSR-S)	0.927	0.936
	Ours (GSR-X)	0.956	0.967
	Ours (Assessor360)	**0.960**	**0.971**

Table 2. Cross-dataset comparison of blind PVQA methods on the CVIQD [30] and OIQA [6] datasets. The arrow points from the training set to the test set.

Dataset	Method	SRCC	PLCC
OIQA ↓ CVIQD	MC360IQA	0.798	0.842
	GSR-X	0.762	0.841
	Assessor360	0.859	0.893
	Ours (ScanpathVQA)	0.733	0.747
	Ours (Assessor360)	**0.872**	**0.904**
CVIQD ↓ OIQA	MC360IQA	0.288	0.349
	GSR-X	0.695	**0.718**
	Assessor360	0.338	0.467
	Ours (ScanpathVQA)	0.636	0.658
	Ours (Assessor360)	**0.703**	0.715

Table 3. Comparison of different scanpath generators for explaining human perceptual scores.

Dataset	Method	SRCC	PLCC
VRVQW	*Human scanpath*	0.786	0.790
	Random sampling	0.075	0.104
	Heuristic sampling [36]	0.431	0.443
	Xu18 [38]	0.712	0.717
	TRACK [25]	0.745	0.749
	Li23 [16]	0.790	0.794
	Ours	**0.805**	**0.814**
CVIQD	Random sampling	0.632	0.640
	Heuristic sampling [36]	0.868	0.870
	ScanDMM [28]	0.856	0.864
	Li23 [16]	0.814	0.827
	Ours	**0.928**	**0.940**
OIQA	Random sampling	0.514	0.536
	Heuristic sampling [36]	0.861	0.872
	ScanDMM [28]	0.865	0.877
	Li23 [16]	0.793	0.799
	Ours	**0.914**	**0.917**

4.2. Main Results

We compare our blind PVQA method with seven existing models, including NIQE [19], MC360IQA [31], Wen24 [35], ScanpathVQA [35], GSR-S [29], GSR-X [29], and Assessor360 [36]. For image quality models such as MC360IQA [31] and Assessor360 [36], we retrain them on the VRVQW dataset. In the case of MC360IQA, we assign the video-level quality score to each key frame and use their temporally averaged score for testing. Assessor360's scanpath generator is adapted to videos by adjusting the semantic context associated with each key frame.

In-dataset Results. Table 1 shows the Spearman's rank correlation coefficient (SRCC) and PLCC[1] results under the in-dataset setting. It is evident that our learned scanpaths enhance the performance of all quality assessors compared to other scanpath-based methods. When integrated with the quality assessor from Assessor360 (*i.e.*, a modified Swin-B with a temporal analysis module), our proposed method achieves the best results on all three datasets. Furthermore, our scanpath generator can boost a simpler quality assessor (*e.g.*, from ScanpathVQA [35] with approximately 11 million parameters) to reach performance levels similar to those of a more sophisticated quality assessor coupled with a weaker scanpath generator (*e.g.*, GSR-S [29] with approximately 112 million parameters). Methods that overlook human viewing patterns, like NIQE [19] and MC360IQA [31], fail to accurately model the human perception of panoramic image and video quality, especially on VRVQW. Additionally, we find that assessing the quality of panoramic videos with authentic distortions tends to be more challenging than for panoramic images with synthetic distortions. This is anticipated because authentic distortions in panoramic videos typically present as a complex blend of various artifacts, localized in space and time.

Cross-dataset Results. Table 2 shows the SRCC and PLCC results under the cross-dataset settings on CVIQD [30] and

[1] As standard practice, we apply a monotonic logistic function to compensate for the nonlinearity in model predictions before computing PLCC.

Table 4. Comparison of different scanpath generators for replicating human viewing patterns using the minimum orthodromic distance (minOD) and maximum temporal correlation (maxTC).

Method	minOD ↓	maxTC ↑
Heuristic sampling [36]	1.325	0.401
Xu18 [38]	1.185	0.637
TRACK [25]	1.067	0.699
Li23 [16]	0.542	0.796
Ours (w/o end-to-end optimization)	0.556	0.781
Ours	**0.536**	**0.805**

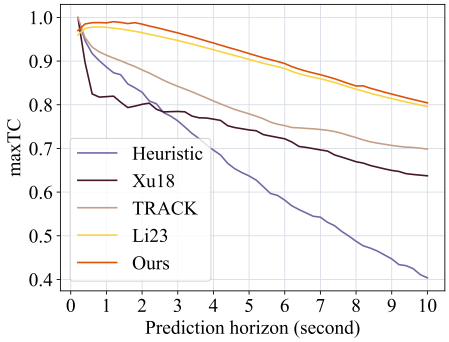

Figure 4. Comparison of different scanpath predictors in terms of maxTC with different prediction horizons.

OIQA [6]. Generally, models trained on the OIQA dataset show better generalizability. This is likely because OIQA encompasses a broader range of distortion types, compared to CVIQD, which only includes the compression artifacts. Assessor360 shows a noticeable performance drop when tested on OIQA, potentially indicative of overfitting. Our scanpath generator is capable of restoring Assessor360's performance, which provides a strong indication of its effectiveness through end-to-end optimization.

4.3. Scanpath Generator Validation

Explaining Human Perceptual Scores. We conduct an apple-to-apple comparison of different scanpath generators in terms of explaining human perceptual scores by fixing the quality assessor to that used in ScanpathVQA. These include random sampling, heuristic sampling [36], Xu18 [38], TRACK [25], ScanDMM [28], Li23 [16], and our method. Table 3 shows the SRCC and PLCC results on the VRVQW, CVIQD and OIQA datasets. It is noteworthy that our scanpath generator outperforms all competing methods across all three datasets, even surpassing the *human-level* performance on VRVQW. The performance of random and heuristic sampling decreases sharply on VRVQW, due to the presence of spatiotemporally localized authentic distortions.

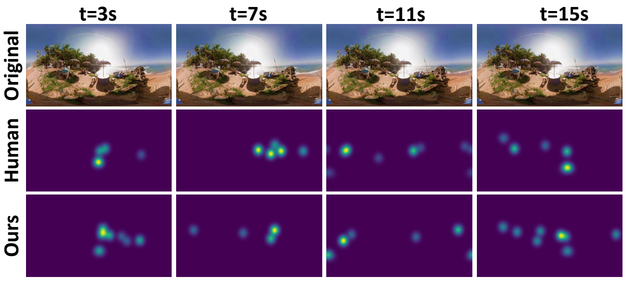

Figure 5. Comparison of saliency maps generated from scanpaths by our method and those by humans.

Table 5. Impact of optimization strategies on blind PVQA. Training protocol: 1) Two-stage w/ fixed scanpaths, 2) Two-stage w/ varied scanpaths, and 3) Three-stage w/ end-to-end optimization.

Protocol	VRVQW		CVIQD		OIQA	
	SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
1	0.769	0.772	0.710	0.780	0.688	0.742
2	0.781	0.785	0.830	0.859	0.798	0.815
3	**0.805**	**0.814**	**0.928**	**0.940**	**0.914**	**0.917**

Replicating Human Viewing Patterns. We also test different scanpath generators in terms of replicating human viewing patterns on the VRVQW dataset by comparing the predicted scanpaths to those of humans. We use two set-to-set evaluation metrics: the minimum orthodromic distance (*i.e.*, minOD) and maximum temporal correlation (*i.e.*, maxTC), as suggested in [16]. Given a set of human scanpaths, $\mathcal{S} = \{\boldsymbol{s}^{(i)}\}_{i=1}^{|\mathcal{S}|}$, the minimum orthodromic distance between $\mathcal{S}$ and the set of predicted scanpaths $\hat{\mathcal{S}} = \{\hat{\boldsymbol{s}}^{(i)}\}_{i=1}^{|\hat{\mathcal{S}}|}$ can be computed by

$$\mathrm{minOD}\left(\mathcal{S}, \hat{\mathcal{S}}\right) = \min_{\boldsymbol{s}\in\mathcal{S},\hat{\boldsymbol{s}}\in\hat{\mathcal{S}}} \mathrm{OD}\left(\boldsymbol{s}, \hat{\boldsymbol{s}}\right), \tag{15}$$

where the orthodromic distance $\mathrm{OD}(\cdot,\cdot)$ is defined as

$$\mathrm{OD}(\boldsymbol{s},\hat{\boldsymbol{s}}) = \frac{1}{T}\sum_{t=0}^{T-1}\arccos\Big(\cos(\phi_t)\cos(\hat{\phi}_t)\cos(\theta_t-\hat{\theta}_t) + \sin(\phi_t)\sin(\hat{\phi}_t)\Big). \tag{16}$$

The maximum temporal correlation between $\mathcal{S}$ and $\hat{\mathcal{S}}$ is defined as

$$\mathrm{maxTC}\left(\mathcal{S}, \hat{\mathcal{S}}\right) = \max_{\boldsymbol{s}\in\mathcal{S},\hat{\boldsymbol{s}}\in\hat{\mathcal{S}}} \mathrm{TC}(\boldsymbol{s},\hat{\boldsymbol{s}}), \tag{17}$$

where the temporal correlation is computed by

$$\mathrm{TC}\left(\boldsymbol{s}^{(i)},\boldsymbol{s}^{(j)}\right) = \frac{1}{2}\left(\mathrm{PLCC}\left(\boldsymbol{\phi}^{(i)},\boldsymbol{\phi}^{(j)}\right) + \mathrm{PLCC}\left(\boldsymbol{\theta}^{(i)},\boldsymbol{\theta}^{(j)}\right)\right). \tag{18}$$

Table 4 presents the minOD and maxTC results, from which we find that our end-to-end optimized scanpath generator delivers the best results, surpassing its independently

Table 6. Impact of visual context on blind PVQA. #Parameters added to the scanpath generator are also shown.

Method	#Parameters	VRVQW		CVIQD		OIQA	
		SRCC	PLCC	SRCC	PLCC	SRCC	PLCC
Ours	1M	0.805	0.814	0.928	0.940	0.914	0.917
Ours w/ visual context	27M	**0.816**	**0.823**	**0.937**	**0.956**	**0.925**	**0.934**

optimized counterpart by a clear margin. The heuristic sampling [36] that depends on the simplified entropy features and equator bias, is inadequate for capturing human viewing patterns, especially for long-term prediction horizons (see Figure 4). Due to the deterministic nature, Xu18 [38] and TRACK [25] fail to accommodate the diversity and uncertainty inherent in human scanpaths, resulting in subpar performance. Incorporating historical video frames as the visual context, Li23 [16] shows performance on par with our method, reinforcing our assertion that visual context informs less about future viewpoints. Figure 5 demonstrates a comparison of the saliency maps derived from scanpaths by our method and those by humans, offering further proof of the close alignment of our scanpath generator and human viewing behaviors.

4.4. Ablation Studies

Impact of Optimization Strategies. We explore three different optimization strategies: 1) a two-stage approach where the pre-trained scanpath generator produces a fixed set of scanpaths for the training of the quality assessor, 2) a similar two-stage approach but supplying a varied set of scanpaths in each epoch of training, and 3) the default three-stage approach that enables end-to-end optimization. From the results in Table 5, we find that the two-stage approach benefits from "data augmentation" with varied scanpaths in each epoch. Our three-stage end-to-end optimization strategy further boosts the accuracy of quality prediction by jointly finetuning both the scanpath generator and quality assessor.

Impact of Visual Context. To assess the impact of visual context on blind PVQA, we enhance our scanpath generator with a video analysis network [16], implemented by a variant of ResNet-50 for frame-level feature extraction and aggregation. We subsequently integrate it with the ScanpathVQA quality assessor for blind PVQA, with the results shown in Table 6. We find that the visual context has a negligible effect on blind PVQA, and thus we exclude it in the generation of scanpaths.

Impact of the Number and Length of Viewport Sequences. We investigate the effects of varying the number N and length L of viewport sequences on blind PVQA. We test N values from $\{5, 10, 15, 20, 50\}$ and L values from $\{4, 7, 15\}$. Figure 6 shows the SRCC results on VRVQW, using the ScanpathVQA quality assessor for prediction. It is clear that $N = 20$ viewport sequences are sufficient for reliable quality assessment, with performance remaining stable with the increase in N. For sequence length, $L = 7$ appears to be a wise choice. Further increasing L does not noticeably affect the performance, but would lead to a considerable rise in computational demand. Conversely, a shorter viewport sequence results in a noticeable drop in performance due to the loss of information from excessive temporal downsampling. It is important to note that these findings are specific to the ScanpathVQA quality assessor and may differ from other assessors.

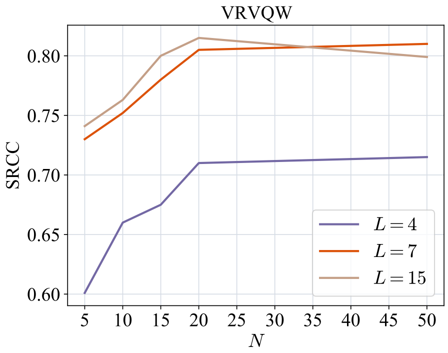

Figure 6. Impact of the number N and length L of viewport sequences on blind PVQA.

5. Conclusion

We have introduced an end-to-end optimized blind PVQA method, consisting of a scanpath generator and a quality assessor. The proposed scanpath generator is differentiable and can be integrated with any planar VQA model, whose effectiveness has been thoroughly validated in supporting blind PVQA and in modeling human viewing patterns. Additionally, we have also devised a three-stage optimization strategy to facilitate training convergence, which aligns with current large-scale optimization practices that involve self-supervised pre-training followed by supervised finetuning, including initial warmup phases [10].

6. Acknowledgement

This work was supported in part by the National Natural Science Foundation of China (62071407 and 62102339), the Shenzhen Science and Technology Program (RCBS20221008093121052), the Research Grants Council of Hong Kong (ECS 2121382, ECS 27212822, and GRF 17201822), and the CCF-Tencent Rhino Bird Fund (9239061).

References

[1] Kiam Heong Ang, Gregory Chong, and Yun Li. PID control system analysis, design, and technology. *IEEE Transactions on Control Systems Technology*, 13(4):559–576, 2005. 4

[2] Christopher M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006. 4

[3] Fang-Yi Chao, Cagri Ozcinar, and Aljosa Smolic. Transformer-based long-term viewport prediction in 360° video: Scanpath is all you need. In *IEEE International Workshop on Multimedia Signal Processing*, pages 1–6, 2021. 2

[4] Sijia Chen, Yingxue Zhang, Yiming Li, Zhenzhong Chen, and Zhou Wang. Spherical structural similarity index for objective omnidirectional video quality assessment. In *IEEE International Conference on Multimedia and Expo*, pages 1–6, 2018. 1, 2

[5] Taco S. Cohen, Mario Geiger, Jonas Köhler, and Max Welling. Spherical CNNs. In *International Conference on Learning Representations*, pages 1–15, 2018. 2

[6] Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Yucheng Zhu, Yi Fang, and Xiaokang Yang. Perceptual quality assessment of omnidirectional images. In *IEEE International Symposium on Circuits and Systems*, pages 1–5, 2018. 2, 5, 6, 7

[7] Jun Fu, Chen Hou, Wei Zhou, Jiahua Xu, and Zhibo Chen. Adaptive hypergraph convolutional network for no-reference 360-degree image quality assessment. In *ACM International Conference on Multimedia*, pages 961–969, 2022. 1, 2

[8] Emil J. Gumbel. *Statistical Theory of Extreme Values and Some Practical Applications*. US Government Printing Office, 1948. 4

[9] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 5

[10] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 16000–16009, 2022. 8

[11] Max Jaderberg, Karen Simonyan, Andrew Zisserman, and Koray Kavukcuoglu. Spatial transformer networks. In *Advances in Neural Information Processing Systems*, pages 2017–2025, 2015. 2, 5

[12] Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with Gumbel-Softmax. In *International Conference on Learning Representations*, pages 1–12, 2017. 2, 4

[13] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *International Conference on Learning Representations*, pages 1–14, 2014. 2, 4

[14] Chen Li, Mai Xu, Lai Jiang, Shanyi Zhang, and Xiaoming Tao. Viewport proposal CNN for 360° video quality assessment. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 10169–10178, 2019. 1, 2

[15] Chenge Li, Weixi Zhang, Yong Liu, and Yao Wang. Very long term field of view prediction for 360-degree video streaming. In *IEEE Conference on Multimedia Information Processing and Retrieval*, pages 297–302, 2019. 2

[16] Mu Li, Kanglong Fan, and Kede Ma. Scanpath prediction in panoramic videos via expected code length minimization. *arXiv preprint arXiv:2305.02536*, 2023. 2, 4, 6, 7, 8

[17] Ze Liu, Yutong Lin, Yue Cao, Han Hu, Yixuan Wei, Zheng Zhang, Stephen Lin, and Baining Guo. Swin Transformer: Hierarchical vision Transformer using shifted windows. In *IEEE International Conference on Computer Vision*, pages 10012–10022, 2021. 5

[18] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video Swin Transformer. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 3202–3211, 2022. 5

[19] Anish Mittal, Rajiv Soundararajan, and Alan C. Bovik. Making a "completely blind" image quality analyzer. *IEEE Signal Processing Letters*, 20(3):209–212, 2012. 6

[20] King-To Ng, Shing-Chow Chan, and Heung-Yeung Shum. Data compression and transmission aspects of panoramic videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 15(1):82–95, 2005. 1

[21] Anh Nguyen, Zhisheng Yan, and Klara Nahrstedt. Your attention is unique: Detecting 360-degree video saliency in head-mounted display for head movement prediction. In *ACM International Conference on Multimedia*, pages 1190–1198, 2018. 2

[22] Bolin Ni, Houwen Peng, Minghao Chen, Songyang Zhang, Gaofeng Meng, Jianlong Fu, Shiming Xiang, and Haibin Ling. Expanding language-image pretrained models for general video recognition. In *European Conference on Computer Vision*, pages 1–18, 2022. 5

[23] David Noton and Lawrence Stark. Scanpaths in saccadic eye movements while viewing and recognizing patterns. *Vision Research*, 11(9):929–942, 1971. 1

[24] David Noton and Lawrence Stark. Scanpaths in eye movements during pattern perception. *Science*, 171(3968):308–311, 1971. 1

[25] Miguel Fabián Romero Rondón, Lucile Sassatelli, Ramón Aparicio-Pardo, and Frédéric Precioso. TRACK: A new method from a re-examination of deep architectures for head motion prediction in 360° videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(9):5681–5699, 2022. 2, 6, 7, 8

[26] Stephen P. Smith. Differentiation of the Cholesky algorithm. *Journal of Computational and Graphical Statistics*, 4(2):134–147, 1995. 4

[27] Xiangjie Sui, Kede Ma, Yiru Yao, and Yuming Fang. Perceptual quality assessment of omnidirectional images as moving camera videos. *IEEE Transactions on Visualization and Computer Graphics*, 28(8):3022–3034, 2021. 1, 2

[28] Xiangjie Sui, Yuming Fang, Hanwei Zhu, Shiqi Wang, and Zhou Wang. ScanDMM: A deep markov model of scanpath prediction for 360° images. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 6989–6999, 2023. 2, 6, 7

[29] Xiangjie Sui, Hanwei Zhu, Xuelin Liu, Yuming Fang, Shiqi Wang, and Zhou Wang. Perceptual quality assessment of 360° images based on generative scanpath representation. *arXiv preprint arXiv:2309.03472*, 2023. 1, 2, 5, 6

[30] Wei Sun, Ke Gu, Siwei Ma, Wenhan Zhu, Ning Liu, and Guangtao Zhai. A large-scale compressed 360-degree spherical image database: From subjective quality evaluation to objective model comparison. In *IEEE International Workshop on Multimedia Signal Processing*, pages 1–6, 2018. 2, 5, 6

[31] Wei Sun, Xiongkuo Min, Guangtao Zhai, Ke Gu, Huiyu Duan, and Siwei Ma. MC360IQA: A multi-channel CNN for blind 360-degree image quality assessment. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):64–77, 2020. 6

[32] 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. 1, 2

[33] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 6000–6010, 2017. 2

[34] Xiaolong Wang, Ross Girshick, Abhinav Gupta, and Kaiming He. Non-local neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 7794–7803, 2018. 2

[35] Wen Wen, Mu Li, Yiru Yao, Xiangjie Sui, Yabin Zhang, Long Lan, Yuming Fang, and Kede Ma. Perceptual quality assessment of virtual reality videos in the wild. *IEEE Transactions on Circuits and Systems for Video Technology*, to appear, 2024. 1, 2, 5, 6

[36] Tianhe Wu, Shuwei Shi, Haoming Cai, Mingdeng Cao, Jing Xiao, Yinqiang Zheng, and Yujiu Yang. Assessor360: Multi-sequence network for blind omnidirectional image quality assessment. In *Advances in Neural Information Processing Systems*, pages 1–22, 2023. 1, 2, 5, 6, 7, 8

[37] Jiahua Xu, Wei Zhou, and Zhibo Chen. Blind omnidirectional image quality assessment with viewport oriented graph convolutional networks. *IEEE Transactions on Circuits and Systems for Video Technology*, 31(5):1724–1737, 2020. 1, 2

[38] 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. 2, 6, 7, 8

[39] Jiachen Yang, Tianlin Liu, Bin Jiang, Wen Lu, and Qinggang Meng. Panoramic video quality assessment based on non-local spherical CNN. *IEEE Transactions on Multimedia*, 23: 797–809, 2021. 1, 2

[40] 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, 2015. 1, 2

[41] Vladyslav Zakharchenko, Kwang Pyo Choi, and Jeong Hoon Park. Quality metric for spherical panoramic video. In *SPIE Optics and Photonics for Information Processing X*, pages 57–65, 2016. 1, 2