Learned Scanpaths Aid Blind Panoramic Video Quality Assessment
Supplementary Material

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

1. Details of the Density Estimation Network

The architecture of the density estimation network is depicted in Figure S1. Of particular interest is the masked computation [S3] used to process the causal path. We refer the readers to the code implementation at https://github.com/kalofan/AutoScanpathQA for detailed parameter configurations.

2. Relative uv Coordinate System

Figure S2 compares different coordinate systems. The transformation $(u, v) = \mathbf{\Psi}_t(\phi, \theta)$ that maps the viewpoint (ϕ, θ) to the (u, v) coordinates relative to the t-th viewport centered at (ϕ_t, θ_t) is broken down into multiple steps. First, (ϕ, θ) is transformed to (x, y, z) in the Cartesian coordinate system:

$$x = r\cos(\phi)\cos(\theta), \tag{S1}$$
$$y = r\cos(\phi)\sin(\theta), \tag{S2}$$
$$z = r\sin(\phi), \tag{S3}$$

where $r = 0.5W_v\cot(0.5\theta_v)$ is the radius of the sphere, determined by the width of the viewport, W_v and the field of view, θ_v. Second, (x, y, z) is rotated with respect to (ϕ_t, θ_t):

$$\begin{pmatrix} x_t \\ y_t \\ z_t \end{pmatrix} = R(\phi_t, \theta_t)^\intercal \times \begin{pmatrix} x \\ y \\ z \end{pmatrix}. \tag{S4}$$

where $R(\phi_t, \theta_t)$ is the rotation matrix defined as the product of two matrices $R_2 \times R_1$:

$$R_1 = \begin{pmatrix} a & -b & 0 \\ b & a & 0 \\ 0 & 0 & 1 \end{pmatrix}, \tag{S5}$$
$$R_2 = \begin{pmatrix} c + (1-c)b^2 & -(1-c)ab & -da \\ -(1-c)ab & c + (1-c)a^2 & -db \\ da & db & c \end{pmatrix}, \tag{S6}$$

where

$$a = \cos(\theta_t), \tag{S7}$$
$$b = \sin(\theta_t), \tag{S8}$$
$$c = \cos(\phi_t), \tag{S9}$$
$$d = \sin(\phi_t). \tag{S10}$$

After rotation, (x_t, y_t, z_t) is transformed to (r, y'_t, z'_t) by projecting it to the plane $x = r$. Last, (r, y'_t, z'_t) is readily represented in the uv coordinate system:

$$u = y'_t + 0.5W_v - 0.5, \tag{S11}$$
$$v = 0.5H_v - z'_t - 0.5, \tag{S12}$$

where H_v is the height of the viewport. We may shift the origin to the viewport center, leading to

$$u = y'_t, \tag{S13}$$
$$v = -z'_t. \tag{S14}$$

3. PID Controller

The PID controller [S1] is a prevalent feedback mechanism that enables continuous modulation of control signals. We adopt the sampling strategy of [S3] and assume a proxy viewer governed by Newton's laws of motion. Initially, the proxy viewer is positioned at the starting point $\hat{\boldsymbol{r}}_{-1} = (0, 0)$ in the uv coordinate system, with an initial speed $\boldsymbol{b}_{-1}$ and acceleration $\boldsymbol{a}_{-1}$. We determine the t-th viewpoint by

$$\hat{\boldsymbol{r}}_t = \hat{\boldsymbol{r}}_{t-1} + \Delta\lambda\boldsymbol{b}_{t-1} + \frac{1}{2}(\Delta\lambda)^2\boldsymbol{a}_{t-1}, t \in \{0, \dots, W-1\}, \tag{S15}$$

where the speed $\boldsymbol{b}_{t-1}$ is updated by

$$\boldsymbol{b}_t = \boldsymbol{b}_{t-1} + \Delta\lambda\boldsymbol{a}_{t-1}. \tag{S16}$$

$\Delta\lambda$ represents the sampling interval, *i.e.*, the inverse of the sampling rate. To update the acceleration $\boldsymbol{a}_{t-1}$, a reference viewpoint $\tilde{\boldsymbol{r}}_t$ for $\hat{\boldsymbol{r}}_t$ is provided by sampling from

*Corresponding author.

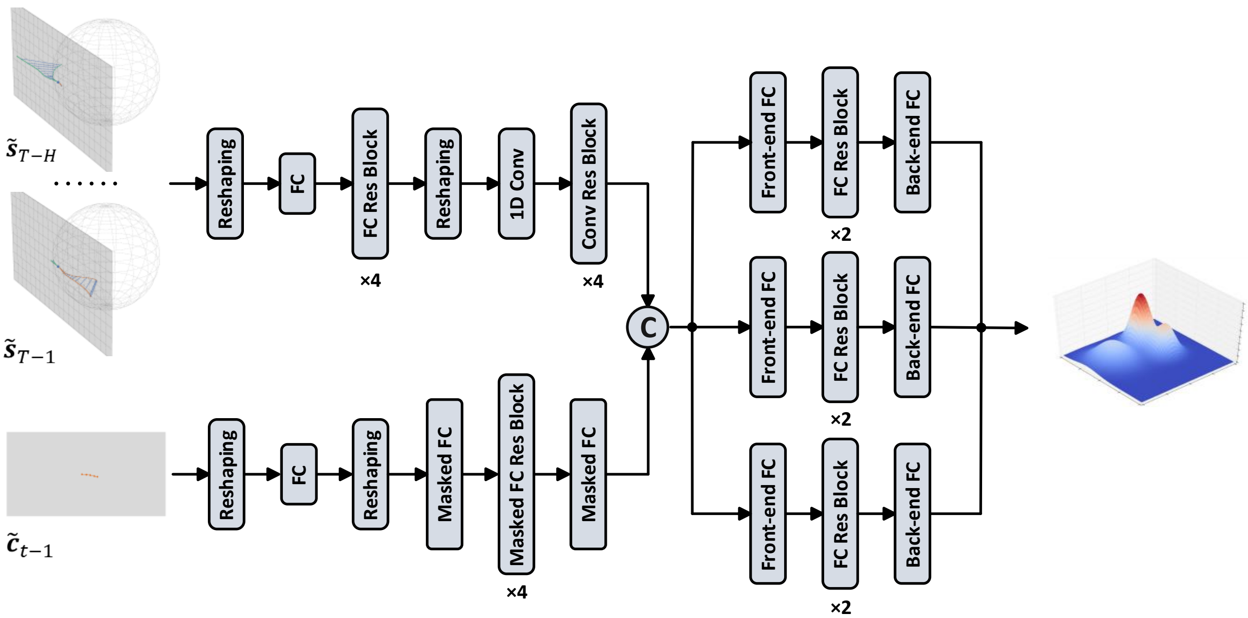

Figure S1. Architecture of the density estimation network.

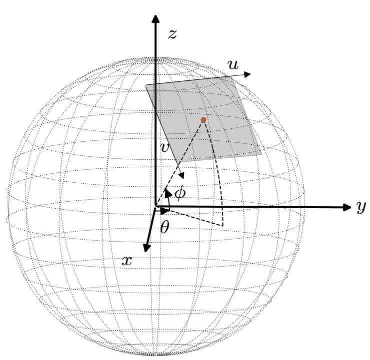

Figure S2. Illustration of different coordinate systems relevant to panoramic signal processing: Spherical Euler coordinates (ϕ, θ), 3D Euclidean coordinates (x, y, z), and uv coordinates (u, v). Image adapted from [S3].

$P\left(\tilde{\boldsymbol{r}}_t \middle| \boldsymbol{s}, \boldsymbol{c}_t\right)$, where $\boldsymbol{c}_t = \{\hat{\boldsymbol{r}}_0, \ldots, \hat{\boldsymbol{r}}_{t-1}\}$. Consequently, an error signal is generated:

$$\boldsymbol{e}_t = \tilde{\boldsymbol{r}}_t - \hat{\boldsymbol{r}}_t, \tag{S17}$$

which is fed to the PID controller for acceleration adjustment:

$$\boldsymbol{a}_t = K_p \boldsymbol{e}_t + K_i \sum_{\tau=0}^{t} \boldsymbol{e}_\lambda + K_d(\boldsymbol{e}_t - \boldsymbol{e}_{t-1}), \tag{S18}$$

where K_p, K_i, and K_d denote the proportional, integral, and derivative gains, respectively. During training, we back-propagate the gradient through the PID controller to the scanpath generator for parameter update.

4. Implementation Details

The first stage of training is carried out by the Adam method [S2] on VRVQW with an initial learning rate of 10^{-4} and a minibatch size of 48. After the 50-th epoch, the learning rate decays by a ratio of 0.1, and we pre-train the scanpath generator for a total of 100 epochs. The parameters for the PID controller are determined using the Ziegler–Nichols method [S5].

For the second and third stages of training, the Adam method is also employed, and the detailed settings of different quality assessors are as follows.

ScanpathVQA. In the second training stage, the quality assessor is trained for 30 epochs, with an initial learning rate of 5×10^{-5}, a decay ratio of 0.95 per 2 epochs, and a batch size of 8. In the third stage of training on VRVQW, the entire method is trained for 5 epochs, with an initial learning rate of 10^{-6}, a decay ratio of 0.1 after the 2-nd epoch, and a batch size of 4. In the third stage of training on CVIQD and OIQA, the method is trained for 5 epochs, with an initial learning rate of 10^{-5}, a decay ratio of 0.9 per epoch, and a batch size of 4.

GSR-S/GSR-X. In the second training stage, we follow the settings described in the original paper [S4]. In the third stage, the entire method is trained for 10 epochs, with an initial learning rate of 10^{-6}, a decay ratio of 0.9 per 2 epochs, and a batch size of 8. The input configuration also follows the original paper [S4].

Assessor360. The settings in the second training stage are identical to those of ScanpathVQA. In the third stage, the entire method is trained for 10 epochs, with an initial learning rate of 10^{-5}, a decay ratio of 0.9 per 2 epochs, and a batch size of 4. Owing to limitations in computer memory, the number of viewport sequences N is reduced from 20 to 15.

References

[S1] 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. 1

[S2] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*, 2015. 2

[S3] Mu Li, Kanglong Fan, and Kede Ma. Scanpath prediction in panoramic videos via expected code length minimization. *arXiv preprint arXiv:2305.02536*, 2023. 1, 2

[S4] 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. 2, 3

[S5] John G. Ziegler and Nathaniel B. Nichols. Optimum settings for automatic controllers. *Transactions of the American Society of Mechanical Engineers*, 64(8):759–765, 1942. 2