Performance Bounds for Computational Models of Visual Saliency in 360 Videos

Aline F. G. Sousa, Clebson I. S. Silva and Ronaldo F. Zampolo

***Abstract*— Estimating performance bounds can improve the comparative analysis of saliency models by exposing their strengths and weaknesses. This paper evaluates two approaches for estimating bounds on immersive videos: Equator Bias and Saliency Sum. For validation, we compare them with outputs from two attention models — Spherical U-Net and DAVE — using three metrics: AUC-Judd, NSS, and CC. Experiments were conducted on 6 videos from the PAVS10K dataset, which includes eye-tracking data from $\sim$20 observers. Saliency Sum achieved the best scores across all metrics, while Equator Bias scored the lowest, indicating that both approaches have significant potential for representing upper and lower performance bounds, respectively.**

***Keywords*— Visual attention modeling, immersive video, performance bounds, performance metrics.**

I. Introduction

Humans are intelligent multisensory creatures that have attentional behavior, that is, they have the ability to detect and focus on specific stimuli in a cluttered environment. In turn, visual attention mechanisms allow to focus on salient regions of a scene, saving important cerebral resources during visual scanning. The phenomenon of visual attention has been studied for over a century and is an interdisciplinary topic that involves several fields of science, including psychophysics, cognitive neuroscience, and computer science [1].

Computational models of visual attention are intended to represent specific aspects of human visual behavior, being generally designed to predict gaze fixation points [2] or to detect salient objects [3].

Currently, we have witnessed the rapid diffusion of 360° videos, also known as immersive or spherical videos. This type of media offers the user the ability to control the viewing angle in a 360° field, a truly immersive experience that enhances viewer engagement. With this rapid spread, the evaluation of visual attention models for immersive videos has initially inherited procedures used for conventional content, which require validation for the new format of media. Among those procedures, there is the definition of lower and upper performance bounds against which the resercher can compare and characterize different saliency models.

Aline F. G. Sousa, Faculty of Computer Engineering, Institute of Geosciences and Engineering, Federal University of the South and Southeast of Pará, Marabá-PA, Brazil (email: alinefarias@unifesspa.edu.br); Clebson I. S. Silva, Signal Processing Laboratory, Postgraduate Program in Electrical Engineering, Institute of Technology, Federal University of Pará, Belém-PA, Brazil (email: clebson.silva@itec.ufpa.br); Ronaldo F. Zampolo, Signal Processing Laboratory, Department of Computer and Telecom Engineering, Institute of Technology, Federal University of Pará, Belém-PA, Brazil (email: zampolo@ufpa.br)

Considering its importance to better evaluate visual attention models, this paper addresses the problem of establishing performance bounds for 360° videos.

The lower bound serves as a criterion for rejecting models, while the upper bound enables a more precise analysis of cost-benefit balance. For example, when introducing a new metric with high computational complexity, one can evaluate whether the gain in terms of similarity to the ground truth justifies the additional cost.

The main contribution of this work is the proposal and evaluation of Equator Bias and Saliency Sum as lower and upper performance bounds, respectively, for visual saliency models in immersive videos. To the best of our knowledge, this is the first time performance bounds for 360° videos have been proposed and discussed.

II. Related works

A. Evaluation of visual saliency models

The performance of a saliency model is assessed by comparing its output with ground-truth data – the latter generally obtained by eye-tracking experiments and whose setups involve a relatively large number of stimuli and subjects. The output of a saliency model and the ground-truth map can be both interpreted as heatmaps indicating the visual relevance of pixels in a depicted scene. Alternatively, they can be also seen as two-dimensional probability density functions. In this sense, the similarity between ground truth and estimated saliency maps would characterize the performance of a visual saliency model, i.e., of how well such a model can predict salient elements of a scene in respect with human viewers.

Next, we review some saliency models with emphasis on the methodology used in their performance evaluation. Starting with models for conventional video, we then move to techniques proposed for immersive media.

Sidaty *et al.* [4] proposed an audiovisual saliency model composed of three maps (spatial, temporal, and auditory), and different fusion strategies to combine such information.

The authors used four performance metrics: area under the ROC curve (AUC), Pearson's correlation coefficient, Kullback-Leibler divergence (KLdiv), and normalized scanpath saliency (NSS).

The Multimodal Saliency model (MMS), proposed by Min *et al.* [5], was developed for 2D videos, whose content has audio information highly associated with the movement of objects. The approach integrates spatial, temporal and sound features. Six metrics are used in the performance evaluation: AUC-Judd, AUC-Borji, shuffled AUC (SAUC), CC, NSS and

earth mover's distance (EMD). The performance of MMS is compared with other visual saliency models like SalGAN [6], SAM-VGG [7], DeepGaze2 [8], among others.

In [9], the Deep Audio-Visual Embedding (DAVE) model is introduced. The proposal uses a deep neural network to integrate visual and auditory information to predict salient regions in 2D videos. The DAVE model was evaluated by five metrics, namely AUC-Judd, SAUC, CC, NSS and similarity (SIM). Besides, the authors adopt the use of performance thresholds to improve their analysis and modulate their expectations. For the lower bound, they consider the mean eye point (MEP), which is based on the general tendency to fixate the center of the screen. And for the upper bound, an approach called human infinite takes subsets of the ground-truth data to estimate the ideal human performance.

For immersive video, Xu et al. [10] proposed a deep learning-based gaze prediction model that incorporates both spatial and temporal information. The model was evaluated using the mean intersection angle error as the main metric. This metric measures the angular difference between the user's gaze prediction and the ground truth.

Cheng et al. [11] proposed a semi-supervised method for predicting saliency in 360° videos using the *cube padding* technique to avoid distortions and discontinuities in the projection of spherical images. The CP360 model is evaluated using AUC-Judd, AUC-Borji, and CC. The results are compared with baselines such as zero-padding.

Nguyen *et al.* [12] proposed a shift from traditional single-viewport saliency models to a new panoramic saliency detection specifically tailored for 360° videos, called PanoSalNet.

In the absence of an eye-tracker the head mounted device, the authors adopted a method similar to Abreu [13], using head orientation as a proxy for gaze fixation. PanoSalNet was evaluated using three performance metrics: sAuC, NSS, and CC. No upper or lower performance bounds were mentioned in the paper.

To summarize, we note that the evaluation of visual saliency models for 360° videos is usually done through comparisons with other models or by variations of the proposed model itself with the support of performance metrics. Strategies and metrics originally developed for 2D saliency models have been adapted to the context of immersive videos, with few studies validading such adaptations.

B. Methodologies for defining performance bounds

Previous studies [14] indicate that, regardless of the observer's task or whether the image features are centralized, fixations tend to accumulate closer to the center of the 2D scenes – an effect called Central Bias (CB).

CB reflects the general tendency for fixations to concentrate in the center of the screen. This bias has been used as a baseline [9] to evaluate the performance of visual attention models in 2D videos. As discussed in [15], any effective saliency model should outperform CB.

While CB is suitable for 2D images, it is considered inadequate for 360° videos or images. In immersive videos, viewers' fixations are usually concentrated around the equator, i.e., exhibiting an EB. This pattern may vary in specific scenarios, but in general, EB remains predominant [16].

An approach for estimating an upper performance bound, also for 2D videos, is the so called human infinite [9]. According to Judd et al., [15] humans are the best predictors of other humans, considering visual tasks. Therefore, it is believed that the saliency map obtained from the fixations of an increasing number of observers (ideally infinite) converges to the optimal predictor of the salient regions of a stimulus.

III. Defining performance bounds for immersive videos

In this section, we detail our proposals for lower and upper performance bounds for evaluating saliency models in immersive videos, namely the Equator Bias and Saliency Sum, respectively.

A. Equator bias

To calculate the version of equatorial bias adopted in this work, we analyzed the fixation data available in the PAVS10K dataset [17]. This particular analysis is based on two histograms showing the distribution of fixations, one histogram in the longitudinal direction and another histogram in the latitudinal direction, covering the entire dataset, i.e. all fixations of all observers from the 67 videos in the dataset. We then calculate the mean and standard deviation of the longitude and latitude distribution, with which we determine a 2D-Gaussian map

$$g(x,y) = k \exp\left\{-\frac{1}{2}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right]\right\}, \quad (1)$$

where x and y denote longitude and latitude in degrees, respectively; σ_x and σ_y are longitudinal and latitudinal standard deviations in degrees, respectively; and k is a normalization factor.

Fixation averages remain relatively close to zero, -2.962 for longitude and 4.669 for latitude. The latter suggests a tendency towards values a little above than the horizontal line. The standard deviation, in turn, indicates greater dispersion in the longitudinal direction than in the latitudinal direction – 76.850 for longitude and 19.816 for latitude – thereby supporting the existence of an Equator Bias.

Figure 1 shows the Equator Bias map generated for the parameters calculated from the PAVS10K dataset.

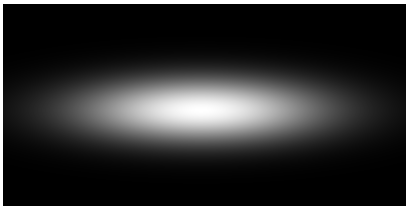


Fig. 1: Two-dimensional Gaussian function generated from gaze statistics of the PAVS10K dataset, representing the Equatorial Bias in immersive videos.

Even though generated by using the reference data, the map in Figure 1 can be considered as ground-truth agnostic from a

practical standpoint, due to the simplicity of its parameters (means and standard deviations). The approach produces a quite rough approximation of gaze patterns, which is independent of specific video structures and can be used as a lower performance bound for immersive videos.

B. Saliency sum

We also propose an approach to estimate an upper performance bound for saliency models in 360° videos, called Saliency Sum. In this approach, the saliency maps of all observers (ground-truth data) are summed across the frames of a video, whose resulting map is then normalized. This process generates a single saliency map per video that is highly dependent on the reference data.

The calculation of a saliency map according to this approach is given by

$$S_{\text{sum}} = \sum_{j=1}^{F} \left(\sum_{k=1}^{N} S_{j,k} \right), \tag{2}$$

where $S_{j,k}$ represents the saliency map (ground truth) in the j-th frame of the k-th observer; N is the number of observers; and F indicates the total number of frames in the video. An example of Saliency Sum map for a video is shown in Figure 2.

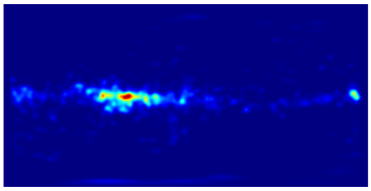


Fig. 2: Example of a saliency map obtained by the Saliency Sum approach applied to a 360° video.

IV. Experimental Procedure

A. Dataset

In this investigation, we use the PAVS10K dataset [17], which contains eye-tracking data from approximately 20 observers for each one of its 67 videos. The results discussed in the next section were obtained using a subset of six videos (Table I), comprising two examples from each of three distinct classes: talking, music, and miscellany. The choice of a stratified subset was motivated by the need to reduce running time while maintaining the significance of the results.

TABLE I: PAVS10K subset main characteristics (FPS: frames per second)

File name	Duration	FPS	#Frames	#Viewers
-J0Q4L68o3xE-1	00'23"	29.97	690	21
-idLVnagjl-s	00'27"	29.97	825	24
-TCUsegqBZ-M	00'25"	25.00	625	20
-gTB1nfK-0Ac	00'29"	29.97	870	24
-RSYbTSTz91g	00'27"	29.97	810	23
-gy4TI-6j5po	00'25"	29.97	750	24

B. Generation of saliency maps

In this study, a saliency map for each frame is calculated from the corresponding eye-tracking data in the PAVS10K, following the standard procedure of convolving a Gaussian function with the frame fixation map.

For 360° videos, the convolution is performed in the equirectangular projection, as the isotropic Gaussian cannot be back-projected onto a sphere [18]. However, it is important to account for the latitude-dependent distortions inherent to the equirectangular projection [19]. Figure 3 gives an example of a frame, and its fixation and saliency maps.

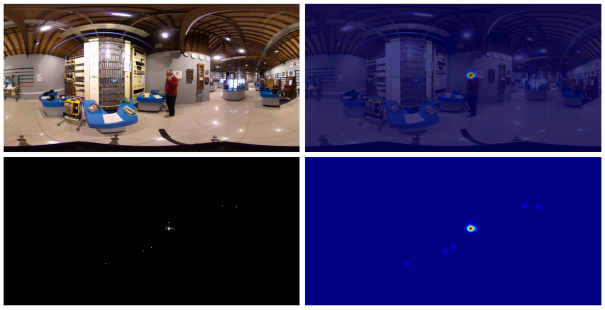


Fig. 3: Examples of frame (top-left), overlay of the frame with its saliency map (top-right), and corresponding fixation (bottom-left) and saliency (bottom-right) maps.

C. Visual saliency models

The proposed performance bounds are tested with the help of two visual saliency models:

1) Deep Audio-Visual Embedding model (DAVE): Tavakoli et al. [9] designed a model to estimate saliency maps for 2D videos. The approach is based on a dual encoder architecture, which uses spatial and temporal attention mechanisms and considers the interaction between visual and audio signals. Its architecture comprises a two-stream 3D convolutional neural network (3D CNN), in which the outputs of the video and audio streams are fused to predict a saliency map of each video frame.

2) Spherical U-Net model: Chuong et al. [20] proposed a spherical convolutional neural network that preserves the perspective of spherical signals, using a spherical kernel that maintains data integrity during convolutions. The model combines contraction and expansion paths with spherical convolutions to predict saliency maps in 360° videos. The inputs include the image of the current frame (3 channels) and the saliency map of the previous frame, totaling 4 channels.

D. Performance metrics

For saliency modeling, performance metrics assess the similarity between the output of a model and the ground truth.

Three metrics were selected for this work:

1) AUC-Judd: This metric determines the area under the Receiver Operating Characteristic (ROC) curve, assessing a model's effectiveness in classifying regions of interest based on human fixations. Scores range from 0 to 1, with values closer to 1 indicating better performance.

2) Normalized scanpath saliency (NSS): This metric measures the agreement between the predicted saliency map and human fixations (ground truth). The NSS is calculated as the mean of the standardized saliency values at the fixation locations [21], as follows:

$$NSS(P, Q^B) = \frac{1}{N} \sum_i \bar{P}_i Q_i^B, \tag{3}$$

with

$$N = \sum_i Q_i^B, \quad \bar{P} = \frac{P - \mu(P)}{\sigma(P)} \tag{4}$$

where P and Q^B are the predicted saliency map and the reference fixation map (ground truth); i refers to the i-th pixel; N is the number of fixations (ground truth); $\mu(\cdot)$ and $\sigma(\cdot)$ denote the mean and standard deviation, respectively.

NSS is a normalized metric, with positive values indicating that human fixations occur in regions of high predicted saliency, i.e., the model is making a good prediction, and negative values indicating fixations in regions of low saliency, meaning the model is failing to correctly predict the areas of interest.

3) Pearson's correlation coefficient (CC): This metric quantifies the linear relationship between the predicted saliency map and the reference saliency map (ground truth). Widely used in visual attention model evaluation, its values range from -1 to 1, where values near 1 or -1 indicate strong positive or negative correlation, respectively, and values near 0 suggest a weak correlation [22]. The CC is defined as:

$$CC\left(P, Q^H\right) = \frac{\text{cov}\left(P, Q^H\right)}{\sigma(P)\sigma\left(Q^H\right)}, \tag{5}$$

where i refers to as the i-th pixel; P is the saliency map representing the model; Q^H denotes the reference saliency map (ground truth); the operators $\text{cov}(\cdot)$ and $\sigma(\cdot)$ refer to the operators that calculate covariance and standard deviation, respectively.

E. Performance assessment

The metrics were calculated individually for each frame, and the overall video score was obtained by averaging the results across the frames. The original spatial resolution was maintained to avoid distortions in the result, although this increases computational complexity.

An important note regarding the equirectangular format, in which the saliency maps are stored: since this projection introduces pronounced geometric distortions in regions closer to the poles, some counterbalance procedure is needed before comparing saliency maps [21].In this sense, we followed the method described in [19], which uses a sinusoidal function to correct oversampled areas across latitudes.

V. Results and Discussion

The Equator Bias and Saliency Sum approaches are provided in Fig. 4, where violin plots illustrate the variation of a given metric across video frames for each proposed approach. Note that, for the tested dataset and metrics, the candidate approaches for providing lower and upper performance bounds are coherent, in the sense that Equator Bias scores are lower than those of the Saliency Sum.

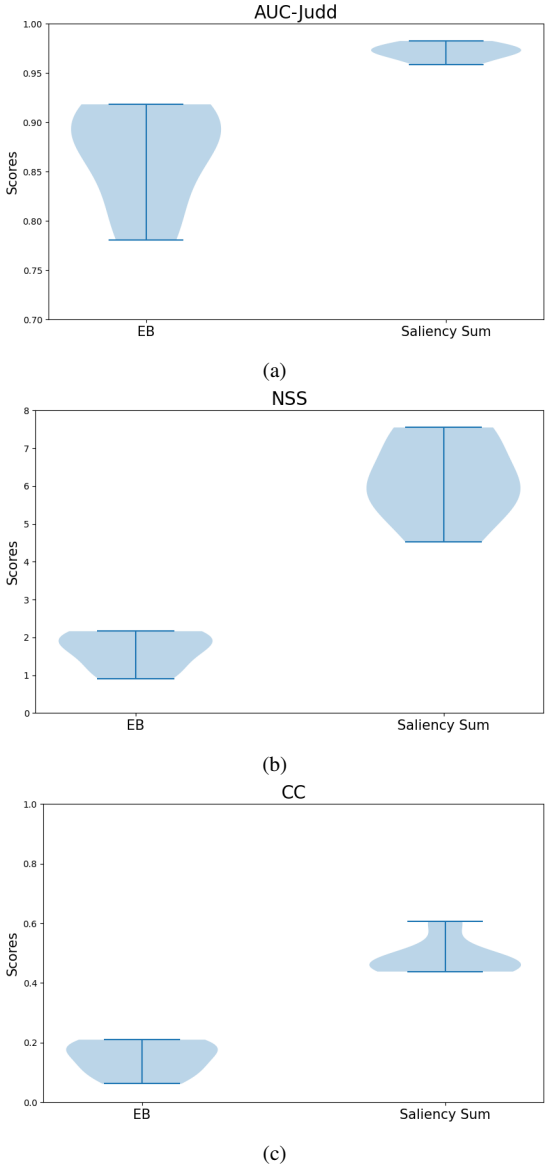


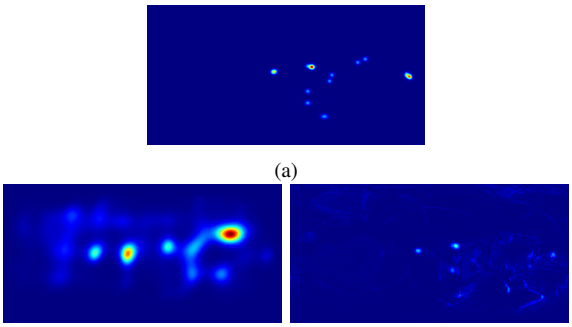
Fig. 4: Scores for metrics: (a) AUC-Judd; (b) NSS; and (c) CC.

Table II confirms such a consistency, presenting the scores averaged across the six videos in the PAVS10K subset for all metrics used. The saliency map generated by the Equator Bias approach showed averaged lower scores across all tested metrics compared to those of the Saliency Sum, which is the desired outcome for candidates representing lower and upper performance bounds, respectively. Table II also presents averaged scores for DAVE and Spherical U-Net saliency models. Both models were tested by using pre-trained versions provided by their original authors and freely available on the internet. The DAVE and Spherical U-Net models performed as expected, with scores within the range given by the Equator Bias and Saliency Sum approaches.

Figure 5 shows qualitative results, comparing the ground-truth saliency map of a frame with the corresponding DAVE and Spherical U-Net estimations.

TABLE II: Performance assessment. The arrow ↑ indicates that the higher the metric value, the better the performance.

Approach/model	Metrics		
	AUC-J ↑	NSS ↑	CC ↑
Saliency Sum	0.9716	6.1479	0.4848
Spherical U-Net	0.9247	5.1508	0.4744
DAVE	0.8747	2.9034	0.2666
Equator Bias	0.8667	1.6611	0.1471



(a)

(b) (c)

Fig. 5: Examples of saliency maps: (a) ground truth; (b) generated by DAVE model; (c) generated by Spherical U-Net model.

VI. Conclusion

This work proposed two approaches for estimating lower and upper performance bounds to evaluate saliency models for immersive videos. The availability of such bounds can enhance the assessment process of saliency models. On one hand, models with performance scores falling below the lower bound should be rejected. On the other hand, competing models with scores close to the upper bound can be considered equivalent and near-optimal solutions.

In a preliminary evaluation with a reduced dataset and selected group of performance metrics, Equator Bias and Saliency Sum approches generated scores consistent to lower and upper bound estimators, respectively. The Equator Bias approach generates a saliency map that, from a practical standpoint, is agnostic to specific video content structures. This approach is grounded in the general visual behavior observed from eye-tracking data, where gaze points tend to concentrate slightly above the horizontal line. The Saliency Sum, in turn, depends on reference data and generates a sort of averaged version of ground-truth saliency maps for each test video.

Despite our promising results, further investigation is necessary to validate the proposed approaches. This includes utilizing an expanded dataset—such as the complete PAVS10K in conjunction with other datasets—and incorporating a broader set of performance metrics.

We are currently investigating the applicability of Human Infinite in the context of 360° videos and plan to present the results in future work.

References

[1] Patrick Le Callet and Ernst Niebur, "Visual attention and applications in multimedia technologies," *Proceedings of the IEEE. Institute of Electrical and Electronics Engineers*, vol. 101, pp. 2058–2067, 09 2013.

[2] Ali Borji and Laurent Itti, "State-of-the-art in visual attention modeling," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 185–207, 2012.

[3] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li, "Salient object detection: A benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[4] Naty Sidaty, Mohamed-Chaker Larabi, and Abdelhakim Saadane, "Toward an audiovisual attention model for multimodal video content," *Neurocomputing*, vol. 259, pp. 94–111, 2017.

[5] Xiongkuo Min, Guangtao Zhai, Jiantao Zhou, Xiao-Ping Zhang, Xiaokang Yang, and Xinping Guan, "A multimodal saliency model for videos with high audio-visual correspondence," *IEEE Transactions on Image Processing*, vol. 29, pp. 3805–3819, 2020.

[6] Junting Pan, Cristian Canton Ferrer, Kevin McGuinness, Noel E O'Connor, Jordi Torres, Elisa Sayrol, and Xavier Giro-i Nieto, "Salgan: Visual saliency prediction with generative adversarial networks," *arXiv preprint arXiv:1701.01081*, 2017.

[7] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara, "Sam: Pushing the limits of saliency prediction models," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1890–1892.

[8] Matthias Kümmerer, Tom Wallis, and Matthias Bethge, "Deepgaze ii: Predicting fixations from deep features over time and tasks," *Journal of Vision*, vol. 17, no. 10, pp. 1147–1147, 2017.

[9] Hamed Rezazadegan Tavakoli, Ali Borji, Esa Rahtu, and Juho Kannala, "Dave: A deep audio-visual embedding for dynamic saliency prediction," *ArXiv e-prints*, 05 2020.

[10] Yanyu Xu, Yanbing Dong, Junru Wu, Zhengzhong Sun, Zhiru Shi, Jingyi Yu, and Shenghua Gao, "Gaze prediction in dynamic 360 immersive videos," in *proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5333–5342.

[11] 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," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, June 2018.

[12] 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 *Proceedings of the 26th ACM International Conference on Multimedia*, New York, NY, USA, 2018, MM '18, p. 1190–1198, Association for Computing Machinery.

[13] Ana De Abreu, Cagri Ozcinar, and Aljosa Smolic, "Look around you: Saliency maps for omnidirectional images in vr applications," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, 2017, pp. 1–6.

[14] Benjamin Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor bases and image feature distributions," *Journal of vision*, vol. 7, pp. 4.1–17, 02 2007.

[15] Tilke Judd, Frédo Durand, and Antonio Torralba, "A benchmark of computational models of saliency to predict human fixations," 01 2012.

[16] 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*, vol. 24, no. 4, pp. 1633–1642, 2018.

[17] Yi Zhang, Fang-Yi Chao, Wassim Hamidouche, and Olivier Deforges, "Pav-sod: A new task towards panoramic audiovisual saliency detection," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 19, no. 3, pp. 1–26, 2023.

[18] Erwan David, Jesus Gutierrez, Antoine Coutrot, Matthieu Perreira da Silva, and Patrick Le Callet, "A Dataset of Head and Eye Movements for 360° Videos," in *9th ACM Multimedia Systems Conference*, Amsterdam, Netherlands, June 2018.

[19] Erwan David, Jesús Gutiérrez, Melissa Le-Hoa Vo, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet, "The salient360! toolbox: Processing, visualising and comparing gaze data in 3d," in *Proceedings of the 2023 Symposium on Eye Tracking Research and Applications*, New York, NY, USA, 2023, ETRA '23, Association for Computing Machinery.

[20] Chuong H. Vo, Jui-Chiu Chiang, Duy H. Le, Thu T.A. Nguyen, and Tuan V. Pham, "Saliency prediction for 360-degree video," in *2020 5th International Conference on Green Technology and Sustainable Development (GTSD)*, 2020, pp. 442–448.

[21] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand, "What do different evaluation metrics tell us about saliency models?," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2019.

[22] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics," *Proceedings of the IEEE International Conference on Computer Vision*, 12 2013.