Use of Saliency Estimation in Cinematic VR Post-Production to Assist Viewer Guidance

Colm O Fearghail[†], Emin Zerman[†], Sebastian Knorr[‡], Fang-Yi Chao[†] and Aljosa Smolic[†]

[†] *V-SENSE, School of Computer Science and Statistics, Trinity College Dublin, Dublin, Ireland*
[‡] *Ernst Abbe University of Applied Sciences Jena, Germany*

Abstract

One of the challenges facing creators of virtual reality (VR) film is that viewers can choose to view the omnidirectional video content in any direction. Content creators do not have the same level of control on viewers' visual attention as they would on traditional media. This can be alleviated by estimating the visual attention during the creative process using a saliency model, which can provide a probability as to what would draw a viewer's eye. In this study, we analyse both the efficacy of omnidirectional video saliency estimation for creative processes and the potential utility of saliency methods for directors. For this, we use a convolutional neural network-based video saliency model for omnidirectional video. To assist the directors in viewer guidance, we propose a metric that provides a measure of saliency estimation in the intended viewport. We also evaluate the selected saliency model, AVS360, by comparing the output of this saliency model to the actual viewing direction. The results show that the selected saliency model can predict the viewers' visual attention well and the proposed metric can provide useful feedback for content creators regarding possible distractions in the scene.

Keywords: omnidirectional video, 360 degree video, saliency, cinematic VR

1 Introduction

Virtual reality (VR) film, also known as cinematic VR, is a form of VR entertainment that utilises among other formats omnidirectional (also known as 360-degree) video. Visual language in VR is still in development, and currently, the techniques used to relate a narrative to viewers in the form are derived from those of traditional cinema [1]. As the viewer has the freedom to look in any direction of the 360-degree environment that they are present in within the format [2], the director of the content must ensure that the narrative is observed as intended by the viewer and to do so in an immersive manner [3].

One method in which to obtain a probability of how a viewer may view a scene is through the use of computational saliency. Saliency models have been developed to evaluate what attracts the human eye within visual scenes [4]. From psychological studies, it is said that bottom-up and top-down processes take place. Bottom-up being the initial attraction based on the physical properties of the image, then the top-down process begins which relates to the task the viewer has while observing the image [5]. These computational models have also been adapted for use within 360-degree video [6].

Using saliency models to estimate the visual attention for a VR film could help the content creator to attract viewers' attention to areas which they deem important to the understanding of the narrative and gain an understanding of other competing salient areas. To allow this, the saliency models can be integrated into post-production environments, as can be seen in Fig. 1.

This publication has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under the Grant Number 15/RP/27760.

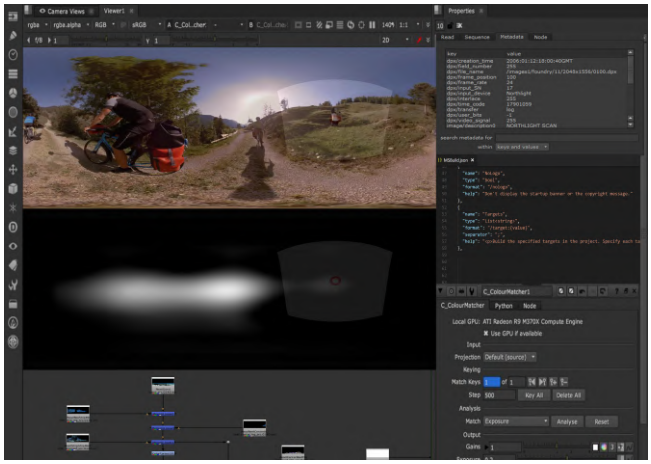


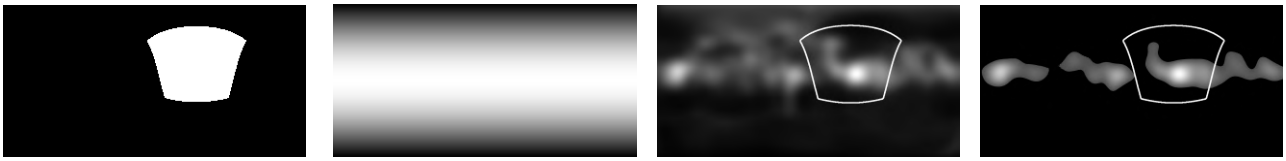
Figure 1: Mock up of a saliency estimator in a post-production environment. The director's cut and corresponding viewport are visible on the RGB image from the film frame and the saliency estimate. The saliency estimate shows that the director's cut viewport is salient, but there is also a region that could compete for viewers' attention. Use of saliency estimation algorithms allows for intervention during post-production.

This paper investigates the effectiveness of using a saliency model in post-production environments, aiming to assist the directors in viewer guidance. With this aim, we build a new metric (i.e., *viewport-based saliency ratio* - VPSR) that can help directors in concentrating the viewers' visual attention and optimising the VR film in the post-production. The proposed VPSR metric can be used with different 360-degree video saliency models. To validate the proposed VPSR metric, we used an omnidirectional video saliency model, i.e., AVS360 [6] developed in our research group for saliency prediction in ODVs (omnidirectional video), and computed the saliency on a VR film database with viewers' visual attention and annotations of director's intended viewing areas, i.e., Director's Cut database [7]. To find out the answer to *"How successful is the selected saliency model in predicting the points that attract visual attention?"*, we first measure the saliency model's output for all the frames of the omnidirectional video. For different contents, the results are then compared to ground truth visual attention to see how well the saliency prediction model performed. Secondly, to answer *"How successful is VPSR in finding frames that need attention?"*, we report the frame-wise results for the proposed viewport-based saliency ratio metric, and we measure how well the director's preferred viewport area is related to the points of saliency within the frame as predicted by the model. Given that AVS360 predicts viewers' attention with high accuracy, the results show the VPSR metric can identify the frames of the video that needs further attention. Our investigation concludes that the use of the proposed metric on saliency estimation methods can identify cases where attention guidance may fail, which can be useful for directors to take appropriate action.

2 Related Work

Among the techniques used by filmmakers in order to communicate their message to the audience are cinematography, mise-en-scène, sound, and editing [8]. In addition to this, various other methods of guiding the viewer within a VR film have been explored. Investigating the methods for guidance, Speicher *et al.* [9] found that giving the viewer an object to follow performed best. Editing from cinematic VR has also formed an area of research [10]. A comprehensive review of papers that have investigated guidance within VR and augmented reality systems can been found in [11].

In order to investigate the ability of techniques derived from traditional cinema within a 360-degree environment, Knorr *et al.* [7] developed a database which included the creator's intended viewing direction at all times throughout the film. This intended viewing direction was given the title of the "Director's Cut" (DC), and this point and corresponding viewport are named as *"DC point"* and *"DC viewport"* throughout this paper. These intended viewing directions were then compared to actual viewing directions of 20 participants to see



(a) Viewport - Ω (b) Weight map - W (c) Saliency map - S (d) Thresholded saliency map - S_{50}

Figure 2: Visualisation of the (a) viewport, (b) weight map, (c) whole saliency map for "Luther" video - Frame #4096 with overimposed viewport, and (d) thresholded saliency map of the same frame for $p = 50$ with overimposed viewport. Please refer to Eqn. 4 for the computation of the VPSR metric.

how viewers consume VR films and how it relates to directors' intentions. Further studies analysed certain elements of the devices used within the films in order to attract the viewers' attention [12] and the styles of cuts, where one scene transitions to another, and their effects on viewer behaviour within the films [13].

To anticipate the user behaviour and estimate viewers' visual attention, saliency estimation methods are developed in image processing and computer vision communities [14]. Due to their spherical nature, omnidirectional images and videos used in immersive systems and VR film are expected to have a different interaction paradigm compared to traditional images and video. How people consume omnidirectional images [15] and video [16] has been explored in the past. Many saliency estimation methods have been developed in the last 20 years [14]; however, more recent advances in the field have been made due to machine learning [17]. An example as to how these models have been used in omnidirectional images can be found in the work of Monroy *et al.* [18] referred to as SalNet360, where the spherical coordinates of the pixels are taken into account. AVS360 [6] which is the saliency model used in this study is a more recent model that caters for omnidirectional video. This model built on work completed in [19]. Development in this area has also included using audio information [20].

To investigate the use of saliency in VR films, we examined the relationship between the SalNet360 saliency estimator and the viewer fixation points at plot points in our previous study [21]. In this earlier study, we focused on plot points in particular and discussed the results for the selected frames. In this paper, differently than in [21], we use AVS360, and we aim to focus on developing a tool that could be used by directors and content creators to identify regions in the scene that could distract viewers from intended viewing areas. To the best of our knowledge, this is the first metric of its kind that will inform directors and content creators.

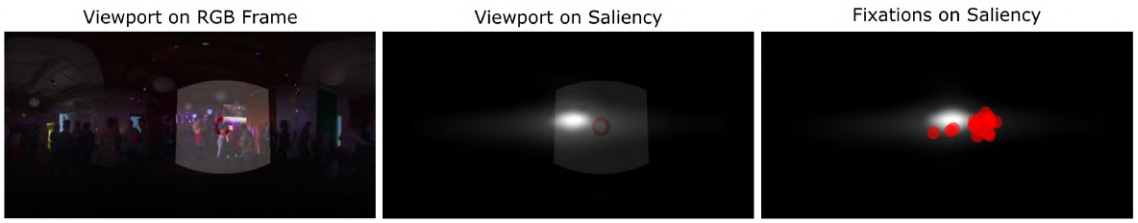
3 Proposed Metric

In this paper, we propose a new metric named *viewport-based saliency ratio* (VPSR) to address the need for a tool that allows directors to optimise their cinematic VR content during post-production. The main goal for this metric is to provide a score to describe the ratio of total probability of estimated saliency within the director's intended viewport. The secondary goal for this metric is to warn the director for possible distractions in the scene that can cause loss of viewers' attention. These distractions then can be avoided using, e.g., virtual effects during post-production.

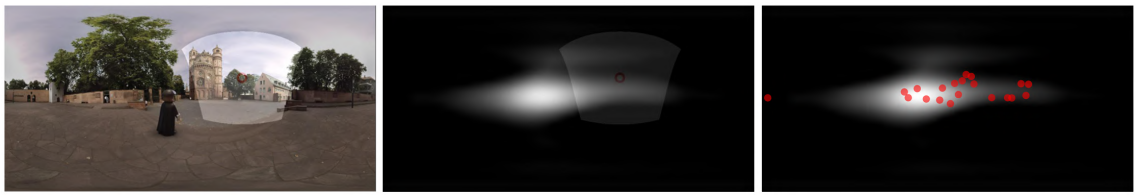
For our VPSR metric, we firstly create a viewport area around the DC point in each frame, which corresponds to the viewport area of the HMD (Head Mounted Display) used in the database study (i.e., an Oculus Rift CV1 headset). Following this, a viewport-based saliency ratio is calculated as follows:

$$\mathrm{VPSR}(S,\Omega) = \frac{\sum_{u,v\in\Omega} S(u,v)W(u,v)}{\sum_{u=1}^{M}\sum_{v=1}^{N} S(u,v)W(u,v)} \tag{1}$$

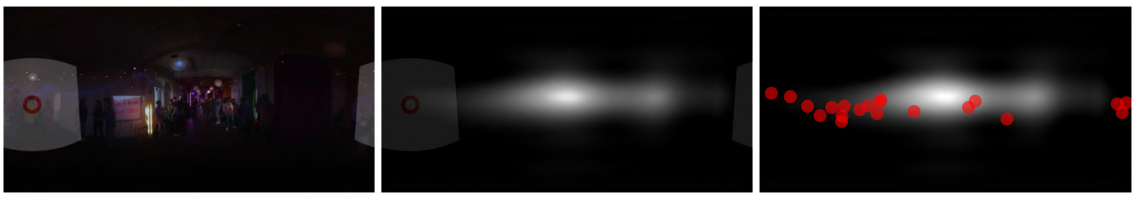
where u and v are the horizontal and vertical pixel locations of an omnidirectional frame of $M \times N$ spatial resolution, Ω is the viewport area as described above (see Fig. 2.(a)), $S(u,v)$ is the saliency map value at (u,v) location, and W is the spherical weighting map. In equirectangular projection, the originally spherical content is increasingly distorted (stretched) along the vertical direction, as we know from geographical maps. The



(a) "Jaunt" - Frame #3680 - VPSR: 0.677 - VPSR_{50}: 0.909



(b) "Luther" - Frame #4690 - VPSR: 0.216 - VPSR_{50}: 0.060



(c) "Jaunt" - Frame #0848 - VPSR: 0.018 - VPSR_{50}: 0

Figure 3: Visualisation of viewers' fixations and viewport corresponding to director's intention from the Director's Cut database [7]. From left to right, *(left)* the RGB film frame with DC point (red circle) and corresponding viewport (white overlay), *(middle)* saliency map with DC point (red circle) and corresponding viewport (white overlay), and *(right)* saliency map with fixation locations (red circles) for two video contents: *(a,c)* "Jaunt" and *(b)* "Luther".

weight map W counters this effect by accounting for spherical distortions, and it gives all parts of the image appropriate contribution to the metric. Here we use the same map as WS-PSNR [22] model, see Fig. 2.(b).

Looking at the distributions of the saliency values with the viewport in Fig. 2.(c) and the fixation distributions in Fig. 3, we notice that the lower saliency values might not attract a lot of viewer fixation. Therefore, we try to generalize the VPSR metric using the saliency values with highest probability. For this, we first compute the histogram of the saliency map S and divide the saliency values into B different bins of $h_i(S)$, where $i \in [1, B]$. Then, taking the highest probability values into account first, we consider the p% of the total saliency and find a threshold value τ for this p as follows:

$$\tau_p = \arg\min_{\tau} \left| \sum_{i=\tau}^{B} h_i(S) - \frac{p}{100} \sum_{i=1}^{B} h_i(S) \right| \tag{2}$$

where B is taken as $B = 256$ for this study. Afterward, using this threshold value, we compute the corresponding thresholded saliency map S_p as follows (see Fig. 2.(d)):

$$S_p(u,v) = \begin{cases} S(u,v), & \text{if } S(u,v) \geq \tau_p \\ 0, & \text{if } S(u,v) < \tau_p \end{cases} \tag{3}$$

This ensures that we always start considering the high probability values. In the last step, we compute VPSR_p as below:

$$\text{VPSR}_p(S,\Omega) = \frac{\sum_{u,v\in\Omega} S_p(u,v) W(u,v)}{\sum_{u=1}^{M}\sum_{v=1}^{N} S_p(u,v) W(u,v)} \tag{4}$$

where $S_p(u,v)$ is the thresholded saliency map value at (u,v) location. We can notice that $S_{100} = S$, and Eqn. 4 is the more generic version of Eqn. 1, i.e., VPSR_{100} =VPSR.

The proposed VPSR metric measures how well the DC viewport captures the estimated saliency compared to the whole saliency map. The metric is bounded between $[0, 1]$, where 0 means no saliency values are under the viewport and 1 means all are under the viewport. Sample VPSR and VPSR_{50} results are given in the captions of Fig. 3.

4 Experiments

Here, we describe the dataset used to analyse the efficacy of omnidirectional video saliency estimation for creative processes, the selected saliency estimation method, i.e., AVS360 [6], and the evaluation metrics used.

4.1 Dataset

In this paper, we use the Director's Cut database [7] to analyse how viewers' fixations relate to the estimated omnidirectional video saliency. This database contains a number of cinematic VR films and includes details from the creators as to where they intended to direct the attention of viewers. For this, creators provided their preferred viewport area throughout the films, using the *Tracker* in the commercial compositing software *Nuke*[1] from The Foundry. The centre of this viewport (i.e., *"DC point"*) is recorded as U and V coordinates, horizontally and vertically. The actual viewing directions were then collected from 20 viewers as they watched the films in a natural manner, by collecting the centre point of viewers' viewports [23]. Further details on this technical process can be read in [7]. Fig. 3 visualises the RGB frames, viewers' fixations, and the estimated saliency maps for three different frames. The first column shows the director's intended viewport overlayed on the RGB frame, the second column shows the director's intended viewport overlayed on the estimated saliency map, and the third column shows participants' fixation points plotted over the saliency map.

For our analysis, we selected four of the films from the Director's Cut database: *"DB"*, *"Jaunt"*, *"Luther"*, and *"Vaude"*. These films had the greatest amount of details as provided by the films' creators, and they also had a range of different lighting and guiding devices used within them.

4.2 Saliency estimation method

To investigate the use of omnidirectional video saliency on VR films and creative processes, we selected the AVS360 model [6] as one of the recent saliency models, the implementation of which is publicly available. This model is composed of two 3D residual networks (ResNets) to encode visual and audio cues. The first one is embedded with a spherical representation technique to extract 360° visual features, and the second one extracts the features of audio using the log mel-spectrogram. While this can take spatial audio into account, the DC database was created with videos using mono sound. The AVS360 model was used as is, without any retraining on the DC database. Interested readers are referred to the original paper [6] for further training details.

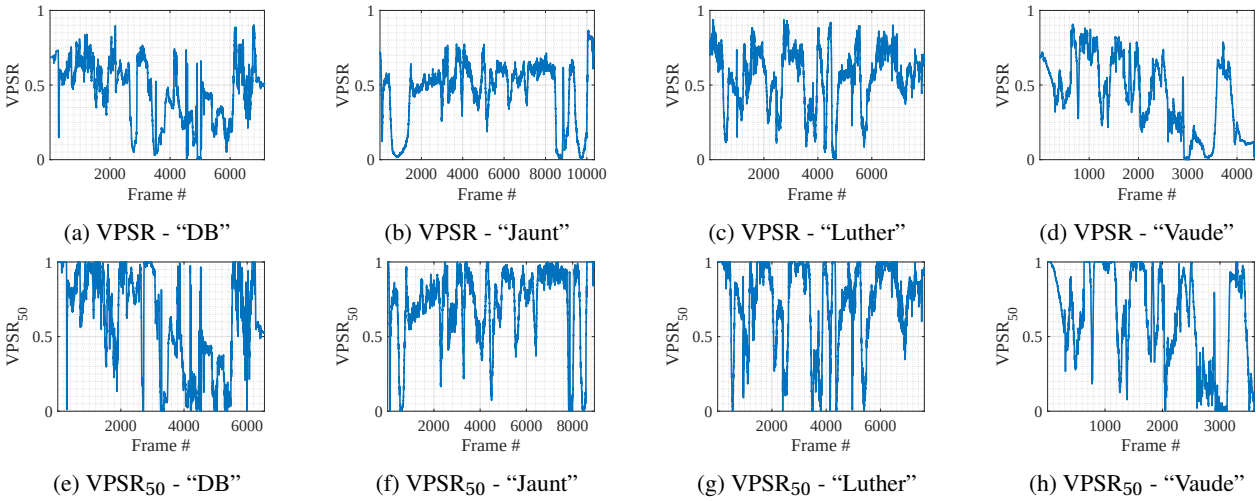
4.3 Evaluation metrics

To evaluate how well the AVS360 model predicts the regions that attract visual attention, we use two saliency evaluation metrics: area under curve (AUC) and normalized scanpath saliency (NSS). Both AUC and NSS are location-based metrics, and they are computed using the ground truth fixation points and estimated saliency map. To compute AUC, the evaluation task was reframed as classification task and the area under the receiver operating characteristic (ROC) curve is computed by finding the true positive and false positive rates. NSS on the other hand first normalises the saliency map (i.e., saliency map is shifted to a mean of zero with standard deviation of one) and estimates the average of the normalised saliency. Additional detail on the metrics used can be found in [24]. To compute these metrics, we used open source implementations for NSS[2] and AUC[3] [25].

[1] https://www.foundry.com/products/nuke

[2] https://sites.google.com/site/saliencyevaluation/evaluation-measures

[3] http://www.saliencytoolbox.net/



(a) VPSR - "DB" (b) VPSR - "Jaunt" (c) VPSR - "Luther" (d) VPSR - "Vaude"

(e) $VPSR_{50}$ - "DB" (f) $VPSR_{50}$ - "Jaunt" (g) $VPSR_{50}$ - "Luther" (h) $VPSR_{50}$ - "Vaude"

Figure 4: Frame-wise evaluation of the relationship between saliency estimation and directors' intention for each video in terms of VPSR.

5 Analysis and Discussion

5.1 Validating the use of AVS360

We first analyse how well AVS360 predicts the ground truth viewing directions. For this, the AUC and NSS metrics are calculated, and the mean AUC and NSS metric results are reported in Table 1.

As we can see from Table 1, both AUC and NSS values support the hypothesis that the selected saliency estimation method, AVS360, can predict the fixation locations well. The AUC score is defined in the range of $[0,1]$, and a high AUC score (e.g., 0.8594 as in Table 1) indicates that the estimated saliency map predicts the distribution of the fixations well. The NSS score shows how large the saliency values correspond to the fixation locations, and having NSS scores ~ 2.5691 means that the saliency values corresponding to fixation locations are 2.5σ away from the mean of the saliency map. That is, the estimated saliency map yields high values at fixation locations. Both of these observations show that the AVS360 model can predict salient regions well. Furthermore, AVS360 can identify locations that might divert visual attention.

5.2 Frame-wise VPSR results

The generic VPSR metric given in Eqn. 4 enables directors to fine-tune the VPSR results by modifying the p value between $[1,100]$. To validate VPSR metric and to show how a change in p affects the results, in this subsection, we provide the frame-wise results for the proposed VPSR metric for two different cases: considering the whole saliency map ($p = 100$) and considering the highest probabilities that sum up to 50% ($p = 50$).

Sample VPSR metric results were provided in the captions of Fig. 3 along with sample frames. These sample results show that a VPSR value of 0.667 corresponds to a very good overlap between the DC viewport and the estimated saliency while a VPSR value of 0.018 indicates poor correspondence. The values are more intuitive for $VPSR_{50}$ as it yields both higher and lower values for these examples. Fig. 4 shows the overall

Table 1: Mean AUC and NSS metric results across all frames comparing saliency maps and viewers' viewing directions.

Film	**"DB"**	**"Jaunt"**	**"Luther"**	**"Vaude"**	**Overall**
$AUC_{Viewers}$	0.8940	0.9264	0.9346	0.8594	0.9036
$NSS_{Viewers}$	1.7367	2.7249	2.7531	2.5691	2.4459

results and allows analysis of how well saliency prediction and directors' intent agree. We can identify dips, which indicate areas that may require intervention to keep viewers' attention. Overall, the VPSR metric was higher for "Luther" compared to other contents. The graphs for VPSR and $VPSR_{50}$ show similar characteristics, while $VPSR_{50}$ has larger swings; therefore, it might provide more intuitive results for the director.

6 Conclusion

In this paper, we proposed a metric that allows directors to optimise their cinematic VR content for viewer guidance. To demonstrate how this metric is capable of yielding useful scores for directors, we used the AVS360 saliency estimation method on an omnidirectional video dataset. We first validated that AVS360 predicts viewers' attention well, and then we presented frame-wise VPSR results. The visual results along with the frame-wise results show that the VPSR metric is indicative of how well the intended viewports could retain viewers' attention.

The results indicate that the AVS360 model and the VPSR metric could form part of a plug-in that will notify the director of regions of possible distractions within the film. The directors will be presented with frame-wise VPSR results as shown in Fig. 4 and they can identify the dips in VPSR values (e.g., dips in visual attention) without checking the saliency estimation results for all the frames manually. With this information the director could then alter the film set during the production or use visual effects (VFX) in post-production accordingly. The VFX option could even be done in an adaptable manner should the viewers' attention stray.

References

[1] K. Dooley, "Storytelling with virtual reality in 360-degrees: a new screen grammar," *Studies in Australasian Cinema*, vol. 11, no. 3, pp. 161–171, 2017.

[2] V. Sitzmann, A. Serrano, A. Pavel, M. Agrawala, D. Gutierrez, B. Masia, and G. 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.

[3] J. Mateer, "Directing for cinematic virtual reality: How the traditional film director's craft applies to immersive environments and notions of presence," *Journal of Media Practice*, vol. 18, no. 1, 2017.

[4] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 1153–1160.

[5] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[6] F.-Y. Chao, C. Ozcinar, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic, "Towards audio-visual saliency prediction for omnidirectional video with spatial audio," in *2020 IEEE International Conference on Visual Communications and Image Processing (VCIP)*. IEEE, 2020, pp. 355–358.

[7] S. Knorr, C. Ozcinar, C. O. Fearghail, and A. Smolic, "Director's cut: A combined dataset for visual attention analysis in cinematic VR content," in *Proceedings of the 15th ACM SIGGRAPH European Conference on Visual Media Production*. ACM, 2018, p. 3.

[8] M. Vosmeer and B. Schouten, "Project orpheus a research study into 360° cinematic VR," in *Proceedings of the 2017 ACM International Conference on Interactive Experiences for TV and Online Video*, 2017.

[9] M. Speicher, C. Rosenberg, D. Degraen, F. Daiber, and A. Krúger, "Exploring visual guidance in 360-degree videos," in *Proceedings of the 2019 ACM International Conference on Interactive Experiences for TV and Online Video*. ACM, 2019, pp. 1–12.

[10] A. Serrano, V. Sitzmann, J. Ruiz-Borau, G. Wetzstein, D. Gutierrez, and B. Masia, "Movie editing and cognitive event segmentation in virtual reality video," *ACM Transactions on Graphics (TOG)*, vol. 36, no. 4, pp. 1–12, 2017.

[11] S. Rothe, D. Buschek, and H. Hußmann, "Guidance in cinematic virtual reality-taxonomy, research status and challenges," *Multimodal Technologies and Interaction*, vol. 3, no. 1, p. 19, 2019.

[12] C. O. Fearghail, C. Ozcinar, S. Knorr, and A. Smolic, "Director's cut - analysis of aspects of interactive storytelling for VR films," in *International Conference on Interactive Digital Storytelling*. Springer, 2018, pp. 308–322.

[13] C. O. Fearghail, C. Ozcinar, S. Knorr, and A. Smolic, "Director's cut - analysis of VR film cuts for interactive storytelling," in *International Conference on 3D Immersion (IC3D)*. IEEE, 2018.

[14] L. Itti and A. Borji, "Computational models: Bottom-up and top-down aspects," in *The Oxford Handbook of Attention*, A. C. Nober and S. Kastner, Eds. Oxford University Press, 2014.

[15] A. De Abreu, C. Ozcinar, and A. Smolic, "Look around you: Saliency maps for omnidirectional images in VR applications," in *Ninth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2017.

[16] C. Ozcinar and A. Smolic, "Visual attention in omnidirectional video for virtual reality applications," in *Tenth International Conference on Quality of Multimedia Experience (QoMEX)*. IEEE, 2018.

[17] K. Zhang, Z. Chen, and S. Liu, "A spatial-temporal recurrent neural network for video saliency prediction," *IEEE Transactions on Image Processing*, vol. 30, pp. 572–587, 2020.

[18] R. Monroy, S. Lutz, T. Chalasani, and A. Smolic, "SalNet360: Saliency maps for omni-directional images with CNN," *Signal Processing: Image Communication*, vol. 69, pp. 26–34, 2018.

[19] F.-Y. Chao, L. Zhang, W. Hamidouche, and O. Deforges, "Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks," in *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2018, pp. 01–04.

[20] F.-Y. Chao, C. Ozcinar, C. Wang, E. Zerman, L. Zhang, W. Hamidouche, O. Deforges, and A. Smolic, "Audio-visual perception of omnidirectional video for virtual reality applications," in *2020 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*. IEEE, 2020, pp. 1–6.

[21] C. O. Fearghail, S. Knorr, and A. Smolic, "Analysis of intended viewing area vs estimated saliency on narrative plot structures in VR film," in *International Conference on 3D Immersion*, 2019. [Online]. Available: https://v-sense.scss.tcd.ie/?attachment_id=4339

[22] Y. Sun, A. Lu, and L. Yu, "Weighted-to-spherically-uniform quality evaluation for omnidirectional video," *IEEE Signal Processing Letters*, vol. 24, no. 9, pp. 1408–1412, Sept 2017.

[23] C. Ozcinar and A. Smolic, "Visual attention in omnidirectional video for virtual reality applications," in *10th International Conference on Quality of Multimedia Experience (QoMEX)*, Sardinia, Italy, May 2018.

[24] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. 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, 2018.

[25] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural networks*, vol. 19, no. 9, pp. 1395–1407, 2006.