https://doi.org/10.2352/EI.2022.34.11.HVEI-168
This work is licensed under the Creative Commons Attribution 4.0 International License.
To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

SalyPath360: Saliency and Scanpath Prediction Framework for Omnidirectional Images

Mohamed Amine KERKOURI[1], Marouane TLIBA[1], Aladine CHETOUANI[1], Mohamed SAYEH[2]
[1]Laboratoire PRISME, Université d'Orléans, Orléans, FRANCE
[2]University of Oran 1 ORAN, ALGERIA

Abstract

This paper introduces a new framework to predict the visual attention of omnidirectional images. The key setup of our architecture is the simultaneous prediction of the saliency map and a corresponding scanpath for a given stimulus. The framework implements a fully encoder-decoder convolutional neural network augmented by an attention module to generate representative saliency maps. In addition, an auxiliary network is employed to generate probable viewport center fixation points through the $SoftArgMax$ function. The latter allows deriving fixation points from feature maps. To take advantage of the scanpath prediction, an adaptive joint probability distribution model is then applied to construct the final unbiased saliency map by leveraging the encoder decoder-based saliency map and the scanpath-based saliency heatmap. The proposed framework was evaluated in terms of saliency and scanpath prediction, and the results were compared to state-of-the-art methods on Salient360! dataset. The results showed the relevance of our framework and the benefits of such architecture for further omnidirectional visual attention prediction tasks.

1. Introduction

Virtual Reality (VR) applications provide high quality of immersive user experiences. Most of VR applications are in the form of 360 video, whereas the frames are represented under a new format of multimedia content called omnidirectional image. These images cover the whole spherical viewing space ($360^\circ \times 180^\circ$), where the user has the freedom of attending to any direction just by pointing his head to any direction. The viewport of the 360° image is defined by the device-specific viewing angle (typically 120 degrees), which delimits horizontally the scene from the head direction center, called the viewport center. The rendering of 360° viewport of the images is supported by many types of sphere to plane coordinates mapping transformations, EquiRectangular Projection (ERP) is one of the most widely used formats of uniform quality mapping projection [21, 19]. It projects the spherical content to a single high resolution 2D plane, where the longitudinal and latitudinal sphere coordinates are represented on the horizontal and vertical ERP axes, respectively.

Unlike the traditional fixed viewport delivery of 2D content, the immersive experience is delivered using recent technologies such as Head Mounted Display (HMD). They are empowered by the ability to investigate spherical space, enabling them to have the best realistic immersive experience with high consumption of resources. Therefore, the capacity to predict the attended viewport that corresponds to the orientations of the head movements beforehand helps to optimize the delivery process and to provide a higher Quality of Experience (QoE) to the viewer [6].

This can be achieved through the prediction of the human visual attention that reflects the most interesting regions within the field of view of the users. This natural mechanism allows humans to explore complex scenes effortlessly and devotes their limited perceptual resources to the most pertinent subsets of received sensory information [35]. The attractive regions, often called salient regions, are usually represented in a heatmap (i.e saliency map). These map models the distribution of the gaze fixations describing the probability that a pixel is salient. The saliency maps are generated by processing the scanpaths of different viewers, which are defined as a sequence of successive fixation points of the viewer's gaze while exploring the image [25]. Predicting such human behaviors is useful in many applications [20]. Saliency has been also successly employed for predicting the quality assessment of different multimedia content [1, 12, 2, 3, 13].

Unlike conventional 2D images, the users are exposed in omnidirectional images to a larger degree of freedom. Visual attention modeling in 360° content is conducted by predicting the center of most probable attended viewports, which reflect the trajectory of the head movements. Moreover, we assume as in [34] that human attention toward omnidirectional content is governed by some statistical bias, as they tend more to equatorial and frontal regions than others, referred to as equator bias.

Studies on omnidirectional content were pioneered by the work of Bogdanova et al. [7, 8] where the spherical static saliency map was generated by normalizing and merging chromatic, intensity and three cue conspicuities. They also created a motion spherical saliency map through a motion pyramid decomposition. At the last stage, the two resulting maps were fused to produce the spherical frame saliency map. In [17], the authors introduced a Fused Saliency Map (FSM) to predict visual attention on ERP omnidirectional images by adopting the well-known 2D saliency model SALICON [24]. In [26], the authors extended both the 2D Boolean Map Saliency (BMS) [36] and GBVS [23] by incorporating the characteristics of ERP images.

With the high performance of deep neural networks in imaging, various saliency prediction models were introduced. In [31], the authors fine-tuned a 2-Dimensional static saliency model, called PanoSalNet, on ERP frames for the task of head movement prediction. The resulting saliency map is further enhanced by prior statistical bias. In [11], the authors fine-tuned also a 2D static model, called SalGAN, on 360° image dataset using cubemap projection and a new objective function, leveraging the combination of three saliency measures. Instead of using a supervised approach to learn saliency from labeled data.In [15], the authors proposed a novel attention-based architecture that adapts the en-

coded latent vector to the characteristics of omnidirectional images through the extended receptive field. They exploited a Cubic Map Projection (CMP) to improve prediction on polar regions. In [16], the authors proposed a new framework that applies existing 2D saliency models on ERP images without requiring in-depth adaptation of the prediction algorithms. They adopt an adaptive weighted joint probability distribution on different kinds of projection of omnidirectional images.

Scanpath prediction methods are more scarce in the literature, even more so in 360° methods. In [38], the authors proposed a model that uses low level features to produce a saliency map. The resulting map is then binarized and fixation points are generated from the obtained binary map using a clustering method. A graph is constructed from the fixation points. The scanpath is generated by maximizing the sum of transfer probabilities on the graph edges. In [5], the authors proposed a hybrid method, that uses deep neural networks and heuristic methods. They use an encoder-decoder network to generate a static saliency volume, from which scanpaths are generated through stochastic techniques. In [4], the authors proposed a deep neural network, that uses an encoder-decoder network and Long Short Term Memory (LSTM) layers to generate scanpaths, combined with adversarial training by Generative Adversarial Network (GAN) architecture.

In this paper, we introduce a new framework to predict the visual attention of omnidirectional images. The proposed architecture, called *SalyPath360*, allows the simultaneous prediction of a saliency map and a corresponding scanpath for a given stimulus.

The main contributions of this paper are as follows:

- Presenting a Neural Network that predicts saliency map and scanpath for 360° in a simultaneous and parallel manner.
- Unlike existing methods, scanpaths are predicted from the refined internal features of a proven saliency prediction model.
- Improving the saliency prediction by using a joint probability mixture between the saliency map predicted by the network and the saliency map constructed from the predicted scanpath.
- Use the $SoftArgMax$ function to predict head scanpaths and train the network seamlessly.

The rest of this paper is organized as follows: Section 2. provides a detailed description of the proposed approach, while Section 3 compares the performance of our approach against state-of-the-art methods. Finally, we conclude our study and discuss the possibility of future improvements in Section 4.

2. SalyPath360 Framework

In this section, we describe in detail the proposed framework (see Fig. 1) which is mainly composed of an encoder-decoder network augmented by a spatial attention module and an auxiliary network that takes the intermediate features at the bottleneck of the encoder-decoder network to generate the corresponding scanpath. In addition, the primary saliency map predicted by the encoder-decoder architecture and the saliency map derived from the predicted scanpath are combined to generate a more representative saliency map. The architecture of each network as well as the attention module used and the merging process applied are described below.

2.1. Encoder-Decoder Network for Saliency Prediction

The encoder-decoder network is used to predict the primary saliency map of 360° images and generate high level intermediate features exploited by the auxiliary network. Inspired by [15] attention stream, the encoder and decoder are composed of four blocks of convolutional layers, interleaved respectively by a (2×2) max-pooling layer and a (2×2) up-sampling layer. At the bottleneck of our network, an attention module is employed to further refine the intermediate features used by the decoder and auxiliary networks. This module takes the representational feature maps and predicts a single channel heatmap, which captures a global representation covering the integrity of the stimulus. It also expands the receptive field of the encoder from (244×244) to (676×676). This expansion is vital to cover the large 360° images. The resulting heatmap is then used as a filter that refines the feature maps generated by the encoder as follows:

$$X' = \gamma \otimes X \otimes Att_S(X) + X \tag{1}$$

where Att_S is the spatial attention module function. X is the input feature maps given by the encoder. X' is the refined feature maps passed to the decoder and auxiliary networks. γ is a learnable parameter, and $\otimes$ represents the element-wise multiplication.

It is worth noting that other sophisticated layers can be used to improve the performance, especially those developed for fine-grained tasks [29, 30, 28].

2.2. Auxiliary Network for Scanpath Prediction

An auxiliary network is also used to generate a scanpath for a given stimulus by mainly leveraging the encoding ability of the encoder-decoder network. It consists of 3×3 convolutional layers, each activated by a $ReLU$ function. The last layer is composed of 100 features maps (i.e. 100 heatmaps), set in accordance to the number of fixation points per user of the considered dataset [33] (see Section). $Soft-Argmax$ (SAM) [27] is then used to estimate the coordinates of a fixation point from each feature map as follows :

$$SAM(x) = \sum_{i=0}^{W} \sum_{j=0}^{H} \frac{e^{\beta x_{i,j}}}{\sum_{i'=0}^{W} \sum_{j'=0}^{H} e^{\beta x_{i',j'}}} (\frac{i}{W}, \frac{j}{H})^T \tag{2}$$

where i, j, i', j' iterate over pixel coordinates. H, W represent the height and width of the feature map, respectively. x is the input feature map and β is a parameter adjusting the distribution of the softmax output.

As this SAM function is differentiable [27], it permits our model to be trained seamlessly unlike the discrete $Argmax$ function. It also allows a sub-pixel accuracy, avoiding the use of up-sampling layers to increase the size of the feature maps, and thus saving resources.

2.3. Merging Probability Distribution of Unbiased Saliency Map Prediction

Let us consider the predicted saliency map T constructed from encoder-decoder architecture (i.e. primary saliency map), and the saliency map S generated from the predicted fixation points by an adequate Gaussian kernel. We notice that T and S could be combined together to get a joint probability distribution [14], making thus a final representative saliency map straightened by the most probable viewport center positions following :

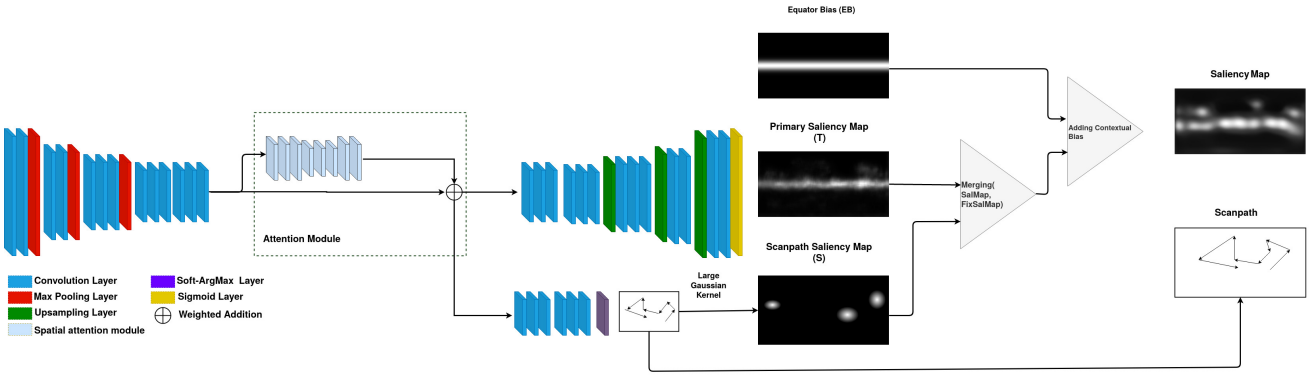

Figure 1. *Proposed framework called SalyPath360.*

$$J(T,S) = \max_{T,S} \times \left(\alpha \left(\frac{T}{\max_{T,S}} \right)^k + (1-\alpha) \left(\frac{S}{\max_{T,S}} \right)^k \right)^{1/k} \quad (3)$$

Where $\max_{T,S}$ is the maximum between the pair (T,S). k is the k-norm for the weighted mean formula and α represents the weight coefficients used to combine T and S distributions. The parameter α was set up throughout an experimental search and was set to 0.7.

For well-tuned α value, the joint probability distribution $J(T,S)$ models a saliency map that integrates spatial and contextual features via the primary saliency map (T) on one side, and the scanpath saliency map (S) generated from the predicted head fixation points on the other. Nevertheless, the variation of head-based movement is about 180° on the Z axis, when considering a spherical referential (X,Y,Z). But in practice, the head variation has a limited range around the equator. Therefore, the computed joint probability distribution $J(T,S)$ should be corrected with another distribution that represents this phenomenon, called the equator bias E. The equator bias E incorporates the most probable head movement bias distribution. With regards to the experiments, we adopted the mean between the previously calculated join probability distribution and the new adjusted one with the equator bias E. The unbiased formula in our context is defined as follows:

$$J^*(T,S) = \left(\frac{1}{2} \times J(T,S) + \frac{1}{2} \times J(T,S) \times L_s(E)) \right) \quad (4)$$

where $L_s(E)$ represents the pixel-wise linear-scaling of the equator bias E.

The applied merging process algorithm for the set of pixels T_{pixels} can be summarized in Algo. 1.

Algorithm 1: `JSalyPath` (IN T,S,E OUT $J^*(T,S)$)

Result: Joint probability distribution $J^*(T,S)$
T Primary saliency map.
S Scanpath-based heatmap.
E Equator bias map.
Compute the normalized linear scaling of bias L_s(E)
L = T_{pixels}
for $(l \in L)$ **do**
  $J(T,S)[l] \leftarrow$
  $\max_{T,S} \times \left(\alpha \left(\frac{T(l)}{\max_{T,S}} \right)^k + (1-\alpha) \left(\frac{S(l)}{\max_{T,S}} \right)^k \right)^{1/k}$;
  $J^*(T,S)[l] \leftarrow$
  $\left(\frac{1}{2} \times J(T,S)[l] + \frac{1}{2} \times J(T,S)[l] \times L_s(E)[l]) \right)$;
end
return $J^*(T,S)$;

2.4. Training

The encoder-decoder and the auxiliary networks were trained through different loss functions. More precisely, the encoder-decoder network based on ATSal [15], which was trained on the Head-Eye movement datasets, was fine-tuned for our task of head movement prediction using the following loss function L_1:

$$L_1 = 0.8 \times KLdiv(y,\hat{y}) + 0.2 \times BCE(y,\hat{y}) - 0.2 \times NSS(y_{fix},\hat{y}) \quad (5)$$

where $KLdiv$ is the Kullback-Leibler Divergence, BCE is the Binary Cross Entropy and NSS is the Normalized Scanpath Saliency. y and $\hat{y}$ represent the ground truth saliency map and the predicted saliency map respectively, while y_{fix} is the ground truth fixation map.

Each term of this loss function was chosen for its own influence on the convergence of the network. Indeed, KLD and BCE functions minimize the distance between the distributions of the output and ground truth, while NSS is a similarity metric for saliency and is used as a bias term, allowing the model to seize the saliency specific representations.

As the auxiliary branch aims to predict the coordinates of fixation points, the task can be seen as a regression problem. In addition, this branch relies upon the feature representations given by the encoder-decoder which is trained by the more complex loss function described-above. As such, this loss (i.e. L_1, see Eq. 5), has an indirect impact on the auxiliary branch during the training step. Therefore, we chose the Mean Squared Error (MSE) as a simple loss function L_2 to train this branch, defined as follows:

$$L_2 = \frac{1}{N} \sum_i (p - \hat{p})^2 \quad (6)$$

where p is the ground truth scanpath and $\hat{p}$ is the predicted scanpath, while N is the number of fixation points of the corresponding scanpaths.

The 2 branches were trained consecutively by first fine-tuning the encoder-decoder network as it has a great influence on

Model	Auc Judd ↑	Auc Borji↑	NSS↑	CC↑	SIM↑	KLD ↓
ATSal[15]	0.8479	0.8121	1.7516	0.6214	0.5748	1.1571
Two-stream [37]	0.7931	0.7564	1.6249	0.5857	0.5857	0.8585
SalyPath360 (Our method)	**0.8610**	**0.8199**	**1.8552**	**0.7194**	**0.6383**	**0.8405**

Table 1. Saliency prediction Comparison. Best results are highlighted in bold.

the accuracy of the auxiliary network. Then, the auxiliary network is trained from scratch while freezing the weights of the encoder in order to not affect the saliency prediction.

3. Experiments

In this section, we evaluate the ability of our model to predict saliency maps and scanpaths. We first describe the used dataset. Then, qualitative and quantitative results are presented and compared with state-of-the-art methods.

3.1. Dataset

Salient360! [33] dataset is one of the most used datasets for predicting the saliency of omnidirectional images. It was proposed as a part of the 2018 Salient360! Challenge and it is composed of 85 omnidirectional images with their corresponding saliency maps and scanpaths. There are about 36 scanpaths per image where each scanpath has 100 fixation points represented by their coordinates and timestamp.

The dataset was split into training-testing sets without any overlap according to the same protocol used in [15]. The training set is composed of 70 images, while the test set is composed of 15 images, representing 82% and 18% of the dataset respectively. This choice of the split was done in accordance with the common practices used in other papers [15] using this dataset. It is worth noting that for a fair comparison, all the compared models were evaluated using the same partition. To the best of our knowledge, Salient360! is the only publicly available dataset that provides the scanpaths in addition to the saliency maps for each 360° image.

3.2. Saliency Prediction

To evaluate the saliency prediction effectiveness of our method, we employed commonly used saliency metrics: *Auc_Judd*, *Auc_Borji*, *NSS* , *CC*, *SIM*, *KLD* [9]. The results are compared to state-of-the-art saliency models: Two-Stream [37] that achieved the best result on Salient360 challenge [22] and ATSal [15] that reached high prediction results Salient360 dataset. It is worth noting that for [37], we used the model published on the leader board of the Salient360! Challenge as Wuhan University. For [15], we used results provided by the authors for their still image model to calculate results.

Table 1 shows the results obtained. Best results are highlighted in bold. As can be seen, the proposed model outperforms all the compared saliency methods, including ATSal. For *Auc_Borji*, we achieve a slight improvement over the state-of-the-art results, while for the other metrics high improvements are noted with a considerable improvement for *NSS* and *CC*. Fig. 2 shows a qualitative comparison between saliency maps predicted using our framework and state-of-the-art models as well the ground truth saliency maps. As can be seen, the saliency maps generated by our framework are closer to the ground-truth.

3.3. Scanpath Prediction

In this section, we evaluate the results obtained by our framework, regarding both the scanpath and the final saliency prediction using common metrics. More precisely, we employed a vector-based metric called *Jarodzka* [18] which compares the similarity between the scanpaths and the hybrid *NSS* [32] metric, which compares the scanpath with the ground truth saliency map. For the former, we applied the code used during the Salient360! challenge [22], while disregarding the temporal element which is not predicted by our framework. We also compare the performance of our framework with state-of-the-art models: PathGan [4] and SaltiNet [5].

Model	Jarodzka ↓	NSS ↑
PathGan[4]	0.1777	-0.1518
SaltiNet[5]	0.2621	0.0834
SalyPath360 (Our method)	**0.1363**	**0.2896**

Table 2. Scanpath prediction comparison. Best results are highlighted in bold.

Table 2 presents the results obtained with the best results are highlighted in bold. As can be seen, our model achieves the best results on *Jarodzka* and *NSS*. PathGan obtains the second best results on *Jarodzka*, while SaltiNet achieves better results on *NSS* metric, outperforming PathGan.

A one way Analysis of Variance (ANOVA) test between groups is also applied to show if the differences between the distributions of the *Jarodzka* values obtained for each compared method are statistically significant. Fig. 4 shows the corresponding boxplots where the middle red mark represents the median value, and the contours of the box are the 25^{th} and 75^{th} percentiles. The extremities of the whiskers correspond to the minimum and maximum values without considering the outliers. The outliers are represented by red crosses and correspond to data points that are further than two or three standard deviations. As can be seen, the distributions are quite different with the smaller median value and data scatter (i.e. standard deviation), followed by those of PathGan. We then computed the p-value between the proposed method and each of the compared methods (i.e. SaltiNet and PathGan). The p-values were lower than the significance level (i.e. 0.05) in both cases, indicating that the differences between the distributions of the compared method are statistically significant. Fig.3 shows a qualitative comparison of the predicted and ground-truth saliency maps as well as those obtained by the encoder-decoder and the auxiliary networks. We also show the corresponding predicted scanpath using SalyPath360. As can be seen, the scanpath predicted through the proposed auxiliary network spans most of the salient regions, while maintaining the bias of equator and frontal regions, but does not visually qualify to be probable. This disparity between the quantitative results obtained through *Jarodzka* metric and the qualitative results shows the limitations of the metric concerning the comparison.

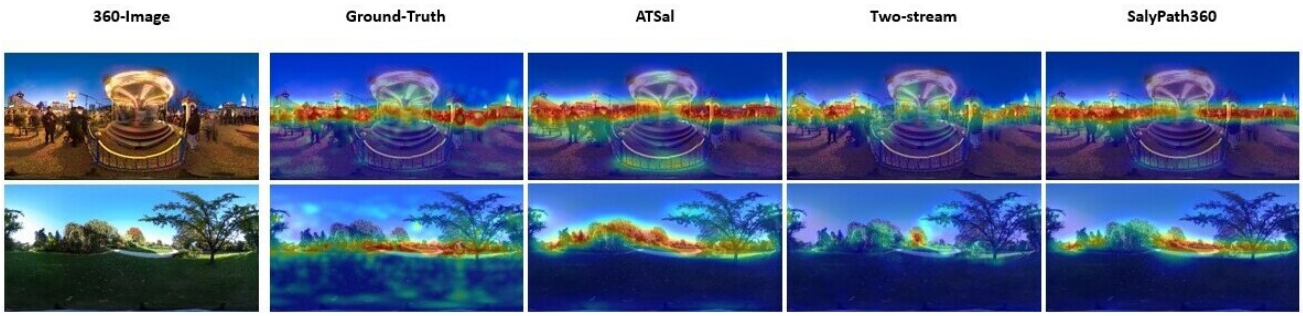

Figure 2. *Qualitative comparison of saliency maps predicted through the proposed SalyPath360 model with state-of-the-art models.*

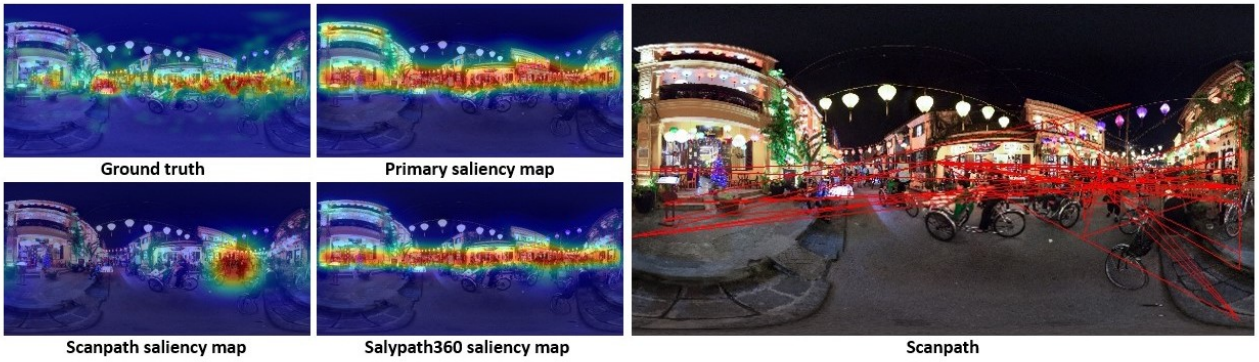

Figure 3. *Qualitative results: On the left part, we show the ground truth saliency map as well as those obtained by the encoder-decoder network, the auxiliary network and the merged map. On the right part, we show the corresponding predicted scanpath using SalyPath360.*

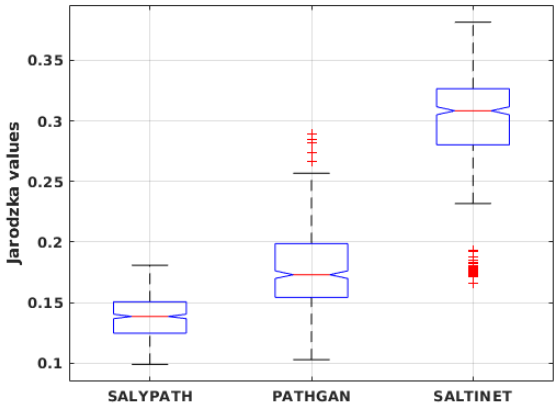

Figure 4. *Results of the one-way Anova test for the $Jarodzka$ values obtained for each compared method.*

3.4. Impact of the joint probability merging

To assess the efficiency of the different components of our framework, we evaluated the saliency maps obtained from the predicted scanpaths, and the saliency maps obtained after the merging without the Equator Bias.

Table 3 displays the results obtained during this study. The maps generated from the scanpath (S) show poor performance for metrics comparing the distributions (i.e. KLD, SIM, CC) while the results for the location-based metrics [10] (i.e. NSS, Auc_Judd, Auc_Borji) show results closer to the comparing models. It is worth noting that the map (S) represents the saliency of a single scanpath. Therefore, the results obtained are satisfactory compared to the ground truth maps aggregating 32 scanpaths. After the merging process (i.e. predicted scanpath generated saliency map and the predicted saliency map), we evaluated the results obtained for J(T,S) in Eq.3. The results on all the metrics showed a slight improvement compared to ATSal and Two-Stream models. While the results after using the Equator Bias show a significant improvement to those after the merging, indicating the beneficial effects of using scanpath prediction and the merging module.

4. Conclusion

In this paper, we introduced a new framework that simultaneously predicts saliency and scanpath for omnidirectional images. The proposed model is composed of an encoder-decoder convolutional neural network for saliency prediction strengthened by an attention module, and an auxiliary network to predict the corresponding scanpath. The latter is then convolved by a Gaussian filter to derive a heatmap. A merging adaptive probability model was finally added at the end of the framework to extract a representative saliency map through the predicted saliency map and the generated scanpath-based heatmap as well as an equator bias map. In our experiments, we trained and tested the framework on the Salient360! dataset. The results were compared with state-of-the-art saliency and scanpath models, showing the effectiveness

Model	Auc Judd ↑	Auc Borji↑	NSS↑	CC↑	SIM↑	KLD ↓
Scanpath Saliency map (S)	0.7746	0.7400	1.4706	0.3629	0.4473	2.0983
Merging map J(T,S) Eq. 3	0.8501	0.8143	1.7547	0.6278	0.5791	1.1105
SalyPath360 (final map)	0.8610	0.8199	1.8552	0.7194	0.6383	0.8405

Table 3. Impact of joint probability module.

of the proposed framework for both tasks. The qualitative results also confirmed the efficiency of our model.

As perspectives, we will further improve the framework by incorporating the temporal dimension, and try to consider the inter-dependence between successive fixation points.

References

[1] Ilyass Abouelaziz, Aladine Chetouani, Mohammed El Hassouni, Longin Jan Latecki, and Hocine Cherifi. Convolutional neural network for blind mesh visual quality assessment using 3d visual saliency. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 3533–3537. IEEE, 2018.

[2] Ilyass Abouelaziz, Aladine Chetouani, Mohammed El Hassouni, Longin Jan Latecki, and Hocine Cherifi. 3d visual saliency and convolutional neural network for blind mesh quality assessment. *Neural Computing and Applications*, 32(21):16589–16603, 2020.

[3] Ilyass Abouelaziz, Aladine Chetouani, Mohammed El Hassouni, Longin Jan Latecki, and Hocine Cherifi. No-reference mesh visual quality assessment via ensemble of convolutional neural networks and compact multi-linear pooling. *Pattern Recognition*, 100:107174, 2020.

[4] Marc Assens, Xavier Giro-i Nieto, Kevin McGuinness, and Noel E O'Connor. Pathgan: Visual scanpath prediction with generative adversarial networks. In *Proceedings of the European Conference on Computer Vision (ECCV) Workshops*, pages 0–0, 2018.

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

[6] Brian Bauman and Patrick Seeling. Spherical image qoe approximations for vision augmentation scenarios. *Multimedia Tools and Applications*, 78(13):18113–18135, 2019.

[7] Iva Bogdanova, Alexandre Bur, Heinz Hügli, and Pierre-André Farine. Dynamic visual attention on the sphere. *Computer Vision and Image Understanding*, 114(1):100–110, 2010.

[8] Iva Bogdanova, Alexandre Bur, Heinz Hügli, and Pierre-André Farine. Dynamic visual attention on the sphere. *Computer Vision and Image Understanding*, 114(1):100–110, 2010.

[9] 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*, 41(3):740–757, 2018.

[10] Z. Bylinskii, Tilke 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*, 41:740–757, 2019.

[11] Fang-Yi Chao, Lu Zhang, Wassim Hamidouche, and Olivier Deforges. Salgan360: Visual saliency prediction on 360 degree images with generative adversarial networks. In *2018 IEEE International Conference on Multimedia & Expo Workshops (ICMEW)*, pages 01–04. IEEE, 2018.

[12] Aladine Chetouani. Convolutional neural network and saliency selection for blind image quality assessment. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 2835–2839. IEEE, 2018.

[13] Aladine Chetouani and Leida Li. On the use of a scanpath predictor and convolutional neural network for blind image quality assessment. *Signal Processing: Image Communication*, 89:115963, 2020.

[14] Roger Cooke et al. *Experts in uncertainty: opinion and subjective probability in science*. Oxford University Press on Demand, 1991.

[15] Yasser Dahou, Marouane Tliba, Kevin McGuinness, and Noel O'Connor. Atsal: An attention based architecture for saliency prediction in 360 videos. *arXiv preprint arXiv:2011.10600*, 2020.

[16] Yasser Dahou, Marouane Tliba, and Mohamed Sayah. 2d-based saliency prediction framework for omnidirectional-360° video. In *Proceedings of the 11th International Conference on Pattern Recognition Systems (ICPRS). accepted and waiting for publication*, 2021.

[17] 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)*, pages 1–6. IEEE, 2017.

[18] Richard Dewhurst, Marcus Nyström, Halszka Jarodzka, Tom Foulsham, Roger Johansson, and Kenneth Holmqvist. It depends on how you look at it: Scanpath comparison in multiple dimensions with multimatch, a vector-based approach. *Behavior research methods*, 44(4):1079–1100, 2012.

[19] Tarek El-Ganainy and Mohamed Hefeeda. Streaming virtual reality content. *arXiv preprint arXiv:1612.08350*, 2016.

[20] Wael Elloumi, Kamel Guissous, Aladine Chetouani, and Sylvie Treuillet. Improving a vision indoor localization system by a saliency-guided detection. In *2014 IEEE Visual Communications and Image Processing Conference*, pages 149–152. IEEE, 2014.

[21] Adriano Gil, Aasim Khurshid, Juliana Postal, and Thiago Figueira. Visual assessment of equirectangular images for virtual reality applications in unity. In *Anais Estendidos da XXXII Conference on Graphics, Patterns and Images*, pages 237–242. SBC, 2019.

[22] Jesús Gutiérrez, Erwan David, Yashas Rai, and Patrick Le Callet. Toolbox and dataset for the development of saliency and scanpath models for omnidirectional/360 still images. *Signal Processing: Image Communication*, 69:35–42, 2018.

[23] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. 2007.

[24] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 262–270, 2015.

[25] Olivier Le Meur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior research methods*, 45(1):251–266, 2013.

[26] Pierre Lebreton and Alexander Raake. Gbvs360, bms360, prosal: Extending existing saliency prediction models from 2d to omnidirectional images. *Signal Processing: Image Communication*, 69:69–78, 2018.

[27] Diogo C Luvizon, Hedi Tabia, and David Picard. Human pose regression by combining indirect part detection and contextual information. *Computers & Graphics*, 85:15–22, 2019.

[28] M.Amine Mahmoudi, Aladine Chetouani, Fatma Boufera, and Hedi

Tabia. Taylor series kernelized layer for fine-grained recognition. In *2021 IEEE International Conference on Image Processing (ICIP)*, pages 1914–1918, 2021.

[29] M Amine Mahmoudi, Aladine Chetouani, Fatma Boufera, and Hedi Tabia. Learnable pooling weights for facial expression recognition. *Pattern Recognition Letters*, 138:644–650, 2020.

[30] M. Amine Mahmoudi, Aladine Chetouani, Fatma Boufera, and Hedi Tabia. Deep kernelized network for fine-grained recognition. In Teddy Mantoro, Minho Lee, Media Anugerah Ayu, Kok Wai Wong, and Achmad Nizar Hidayanto, editors, *Neural Information Processing*, pages 100–111, Cham, 2021. Springer International Publishing.

[31] 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*, pages 1190–1198, 2018.

[32] Robert J Peters, Asha Iyer, Laurent Itti, and Christof Koch. Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18):2397–2416, 2005.

[33] Yashas Rai, Jesús Gutiérrez, and Patrick Le Callet. A dataset of head and eye movements for 360 degree images. In *Proceedings of the 8th ACM on Multimedia Systems Conference*, pages 205–210, 2017.

[34] Mai Xu, Yuhang Song, Jianyi Wang, MingLang Qiao, Liangyu Huo, and Zulin Wang. Predicting head movement in panoramic video: A deep reinforcement learning approach. *IEEE transactions on pattern analysis and machine intelligence*, 41(11):2693–2708, 2018.

[35] Alfred L Yarbus. Saccadic eye movements. In *Eye Movements and Vision*, pages 129–146. Springer, 1967.

[36] Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *Proceedings of the IEEE international conference on computer vision*, pages 153–160, 2013.

[37] Kao Zhang and Zhenzhong Chen. Video saliency prediction based on spatial-temporal two-stream network. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(12):3544–3557, 2018.

[38] Yucheng Zhu, Guangtao Zhai, and Xiongkuo Min. The prediction of head and eye movement for 360 degree images. *Signal Processing: Image Communication*, 69:15–25, 2018.