Simple baselines can fool 360° saliency metrics.

Yasser Abdelaziz Dahou Djilali, Kevin McGuinness and Noel E. O'Connor
Insight Centre for Data Analytics, Dublin City University (DCU)
yasser.dahoudjilali2@mail.dcu.ie

Abstract

Evaluating a model's capacity to predict human fixations in 360° scenes is a challenging task. 360° saliency requires different assumptions compared to 2D as a result of the way the saliency maps are collected and pre-processed to account for the difference in statistical bias (Equator vs Center bias). However, the same classical metrics from the 2D saliency literature are typically used to evaluate 360° models. In this paper, we show that a simple constant predictor, i.e. the average map across Salient360 and Sitzman training sets can fool existing metrics and achieve results on par with specialized models. Thus, we propose a new probabilistic metric based on the independent Bernoullis assumption that is more suited to the 360° saliency task.

1. Introduction

Panoramic images provide users with the ability to explore different regions of the viewing sphere. The average person's head movements (HM) are typically a good prediction of the most probable viewport localized within the sphere, while eye movements (EM) reflect regions-of-interest (RoIs) inside the predicted viewports. Thus, when predicting the most salient pixels for 360° images, it is necessary to predict both HM and EM [25]. Despite much progress in head/eye movements prediction for panoramic scenes in recent years, benchmarking models can be problematic due to the inconsistent behaviour of the metrics used. Many metrics have been adopted or specifically designed for saliency to assess progress and compare models [22, 10, 20, 11, 13]. The AUC set of metrics are the traditional measure, but recently other metrics like Correlation Coeficient (CC), Normalized Scanpath Saliency (NSS), Similarity metric (SIM) and Kulback-Leibleir Divergence (KLD) have become the primary measures used. Please see [1] for a detailed review of these metrics. Salient360! [8], the only available benchmark for 360° saliency, evaluates models based on five metrics directly extended from the 2D saliency literature. Given that models are ranked on the basis of these metrics, many authors have analyzed the metrics both theoretically and experimentally to give recommendations on which is the most appropriate [15, 24, 21]. Other approaches have proposed different loss functions [9]. In this paper we highlight issues with the existing metrics for 360° saliency and propose a more appropriate alternative.

2. Related work

The work reported in [21] performed experiments on Jian Li's human eye-tracking fixations dataset [16], and concluded that sAUC and KLD form a separate cluster from NSS, SIM, and CC. This is due to the fact that KLD assumes saliency maps to be strongly regularized, and sAUC does not to account for the built-in center bias. The study of [6] randomly split the ground truth fixations into two sets to create the reference fixation map and the human fixation map. This human consistency test suggested that NSS and CC capture enough information and KLD is the worst. Based on subjective studies, the authors of [17] argued that human perception is driven by the most salient and high energy regions. Thus, NSS, CC, and SIM are the most consistent with human subjects. AUC reports unsatisfactory results given that it relies solely on the ROC curve without considering the distribution of thresholds. The authors developed a CNN-based metric using the users' data answers as labels. [1] conducted an extensive survey on all metrics, and argued that properties of the inputs affect metrics differently: how the ground truth is defined; whether the prediction takes dataset statistical bias into account; whether the inputs are probabilistic. The authors' main claim is that adapting the properties of metrics for the downstream task can guide metric selection for saliency model evaluation. [13] attempted to solve this issue by calculating the information gain i.e. log-likelihood, the essence of which is to jointly optimize for the scale, the center bias, and spatial blurring as a pre-processing step for all saliency maps to avoid these confounding factors for model comparison. Their approach obtained a new consistent ranking of models. However, it assumes access to all models, which is not practical. Furthermore, the log-density performs suboptimally on most metrics and can still

produce inconsistent rankings [14]. Inspired from Bayesian decision theory, [14] separated the saliency model from the saliency map, where a specific saliency map can then be submitted to a certain metric. Consequently, saliency models should be defined as metric-independent probability densities over possible fixations and subsequently many different metric-dependent saliency maps can be derived from the same density for different error metrics. In this way, saliency models can be meaningfully compared on all metrics at their original scale.

The consensus among the research community is that the metrics measure different things [24, 21, 1], and that it is conceptually impossible to determine a best performing model independent of considering the different metrics. In this paper, we highlight the lack of active research on this topic for 360° visual attention modelling and the different underlying assumptions for 360° data related to the visual attention mechanism [25]. The main contributions of this paper are: (1) We investigate the behaviour of the traditional metrics on 360° saliency datasets, and show that a simple constant predictor (i.e. the average saliency across the training set) can obtain high scores calling into question the practical usefulness of these metrics under the 360° saliency setting. (2) We propose an alternative formulation of the equi-rectangular saliency map as a probabilistic map with the Bernoulli assumption, where each pixel represents the probability of a fixation occurring. (3) We adapt the probabilistic metrics (KLD and JSD) to this Bernoulli assumption and show that these new metrics are a better basis for ranking models.

3. Analysis

3.1. Evaluation setup

The behaviour of metrics and the properties of saliency maps are investigated using the following set of baselines. **The constant predictors** are the average saliency maps across the Salient360! and Sitzman dataset separately. **The equator bias model** is a symmetric Gaussian around the equator; it consists of a degenerate Gaussian with infinite variance in the x-direction and a variance to cover 20% of the equator in the y-direction. This choice is motivated by the statistics of the amount of fixation vs. non-fixation points of an average saliency map. **The random model** gives a uniformly distributed saliency value to each pixel in the saliency map. These baselines are compared against deep saliency models in the literature using the five classical metrics (AUC-J, NSS, CC, SIM, KLD) as shown in Table 1.

Both the equator bias and the random model are image-independent, dataset-independent models. The five metrics are significantly higher for the equator bias model, indicating the importance of capturing the dataset priors. The high number of zeros is detected by the KLD metric as it is sensitive to insufficient regularization. The random model performs poorly across all metrics, caused by the presence of a large amount of false positive/negatives. AUC-J is the least affected by low-valued false positives, and thus reports a relatively higher score **(0.640)** compared to other metrics.

To a large extent, however, the constant predictors (image-dependent, dataset-dependent) succeed in fooling the metrics, and produce scores comparable with deep learning models that were trained end-to-end on these datasets. This has three potential explanations. First, the models may simply be learning to predict the average saliency across the training set. Second, it could be that the datasets are insufficiently varied and there is a high correlation between scenes. The final interpretation relates to the representation of the saliency maps and the definition of the metrics.

Issues with the models. The models' predicted maps look more targeted, and attend to specific regions of the input. The correlation between the models' predictions and the constant predictor is low and suggest that the models are not simply predicting the average training maps (see Figure 1 and Figure 2).

Dataset issue. From [7], Salient360! stimuli are dependent on various considerations, such as shooting environment, amount of foreground objects, distance to objects, presence of people, etc. The eye-gaze data was gathered from 40 subjects per scene, thus, we believe the dataset is well-designed.

Questioning the metrics/saliency formulation. Saliency is extracted from the coordinated motion data of the eyes and head to perform the attention task while moving in the panoramic scene. Measuring the quality of the models predicting the head and eye movements is an open research question. The underlying assumptions for 360° scenes could affect the properties and behaviour of metrics. We argue that directly applying the definition of saliency and its expected input from the 2D context is problematic and will result in some diverging and inconsistent behavior of the metrics.

3.2. Saliency formulation

Definition. The ground truth saliency maps are computed by convolving each fixation or trajectory points (for all observers of one image), defined as:

$$FM_{ij} = \begin{cases} 1 & \text{if location } (i,j) \text{ is a fixation} \\ 0 & \text{otherwise,} \end{cases}$$

with a Gaussian or Kent kernel. The resulting saliency map $P \in [0,1]^{W \times H}$ can be treated as a multivariate Bernoulli distribution where each pixel is Bernoulli distributed, with a probability p to be attended, and $(1-p)$ to be discarded. Probabilistic metrics induce a multinomial distribution on the predictions using a softmax, this assumes that viewers attend to one pixel for all time by dividing by the sum, concentrating

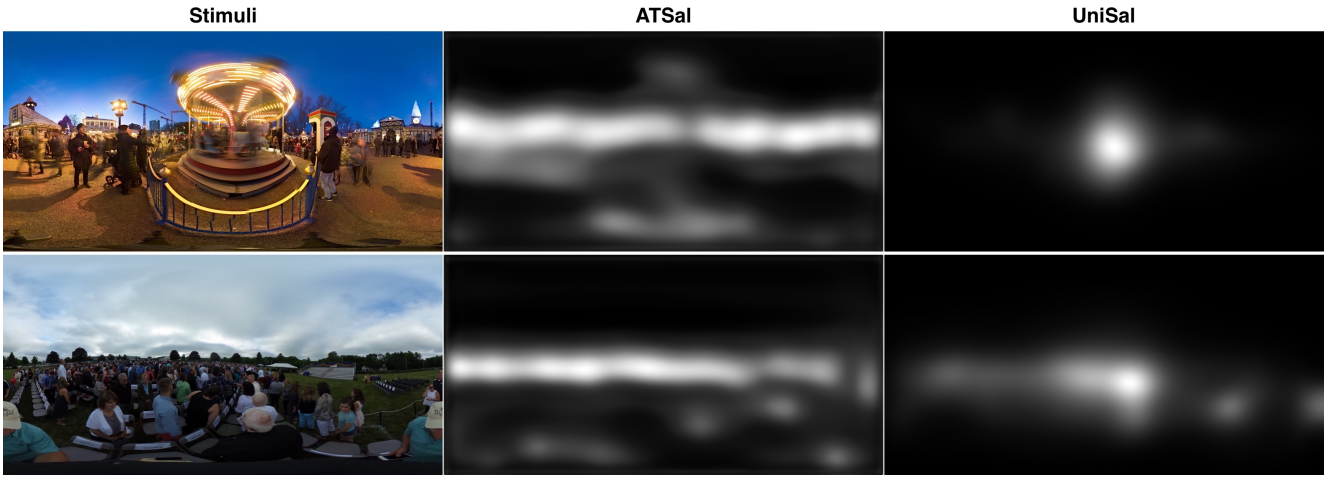

Figure 1. Predicted saliency maps from Salient360! dataset – samples of ATSal [3] and UniSal [4].

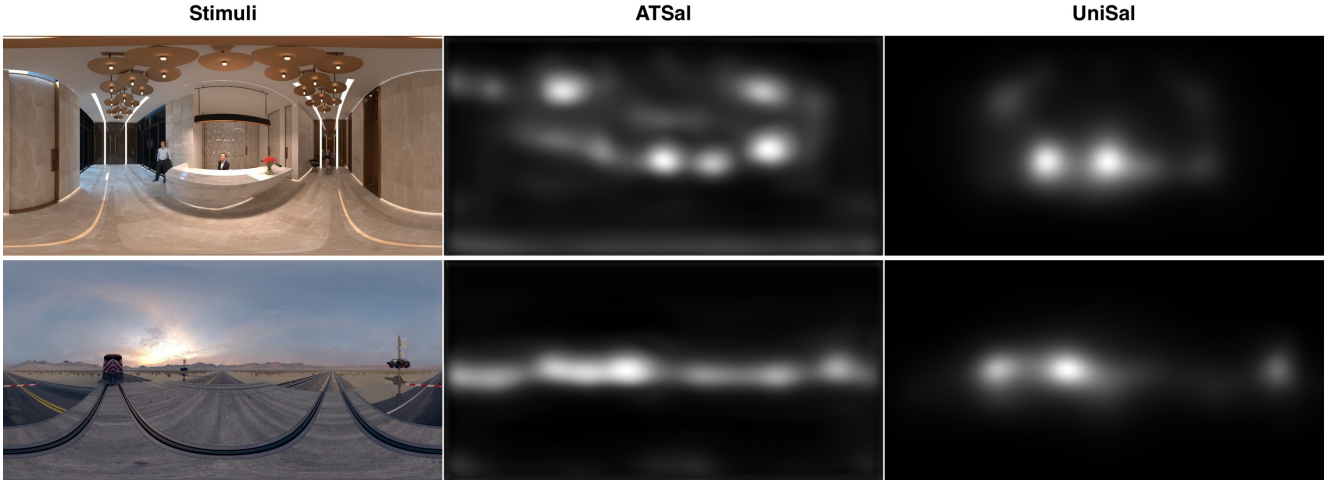

Figure 2. Predicted saliency maps from Sitzman dataset – samples of ATSal [3] and UniSal [4].

the energy on one point. Clearly, however, more than a single pixel may be attended, making it more appropriate to treat each predicted value as independent of the others. Thus, we believe the Bernoulli assumption makes more sense. Multivariate Bernoulli distributions are established using the concept of the Knonecker product from matrix calculus, and provide an alternative to the traditional log-linear models for binary variables [23].

f-divergences used as general (entropic) distance-like functions, the KLD: $P_1 \times P_2 \leftarrow [0, \infty]$ is an oriented statistical distance measured between two densities p and q (i.e. the ground truth and the model predicted saliency maps represent the respective densities). A common symmetrization of the KLD is the Jensen-Shannon Divergence (JSD), also referred as the capacitory discrimination, which is the total KLD to the average distribution $\frac{p+q}{2}$, and is usually applied to densities with arbitrary support. Both KLD and JSD expect the inputs to be a valid probability distribution (i.e. softmax over saliency map pixels), thus, making the assumption of one single point is fixated upon at all times. However, the spherical representation of the panoramic scenes allows viewers to explore many viewports (head movements) – the softmax over the ERP saliency map pixels puts most of the energy on the highly attended viewport by the average observer. We argue that treating saliency pixels as independent Bernoulli events better matches the subjects' behavior, and is less harmful when evaluating models since it compares saliency maps on pixel-wise and doesn't involve any prior regularization. We redefine KLD and JSD as explained in the following.

Bernoulli KL-divergence The formula for KL divergence that is usually used to evaluate saliency maps is based on the assumption that ground truth saliency q is a categorical distribution over the pixels in the image, i.e. $q = (q_1, \ldots, q_N), q_i \geq 0, q^T \mathbf{1} = 1$. Given a softmax prediction

$p = (p_1, \ldots, p_N), p_i \geq 0, p^T \mathbf{1} = 1$ the KL-divergence between the predicted distribution and the ground truth is given by:

$$\mathrm{KL}(q \mid\mid p) = \sum_{i=1}^{N} q_i \log \frac{q_i}{p_i},$$

with appropriate regularization constants added for numerical stability. Here we assume that the ground truth saliency is not normalized to sum to one, but rather represents the probability that each pixel will be attended, independent of the others, i.e. $y_i \sim \mathrm{Bern}(q_i)$ where y_i is a binary RV indication that a fixation occured at pixel i. The per-pixel KL divergence is then given by:

$$\mathrm{KL}(q_i \mid\mid p_i)_{\mathrm{Ber}} = q_i \log \frac{q_i}{p_i} + (1 - q_i) \log \frac{1 - q_i}{1 - p_i},$$

and we propose to minimize the average KL-divergence between the ground truth and predicted saliency map:

$$\mathrm{KL}(q \mid\mid p)_{\mathrm{Ber}} = \frac{1}{N} \sum_{i=1}^{N} \left[q_i \log \frac{q_i}{p_i} + (1 - q_i) \log \frac{1 - q_i}{1 - p_i} \right].$$

Bernoulli Jensen-Shannon divergence The Jensen-Shannon (JS) divergence between two distributions q and p defined on the same alphabet is given by:

$$\mathrm{JS}(q \mid\mid p) = \frac{1}{2}\mathrm{KL}(p \mid\mid m) + \frac{1}{2}\mathrm{KL}(q \mid\mid m),$$

where $m = (p + q)/2$ is the midpoint distribution. Incorporating the Bernoulli KL-divergences of pixel i and simplifying gives:

$$\begin{aligned} \mathrm{JS}(q_i \mid\mid p_i)_{\mathrm{Ber}} = \frac{1}{2} \Bigg[& p_i \log \left(\frac{2p_i}{p_i + q_i} \right) + q_i \log \left(\frac{2q_i}{p_i + q_i} \right) \\ & - (p_i - 1) \log \left(\frac{2(p_i - 1)}{p_i + q_i - 2} \right) \\ & - (q_i - 1) \log \left(\frac{2(q_i - 1)}{p_i + q_i - 2} \right) \Bigg]. \end{aligned}$$

Again, we propose to minimize the average JS-divergence:

$$\mathrm{JS}(q \mid\mid p)_{\mathrm{Ber}} = \frac{1}{N} \sum_{i=1}^{N} \mathrm{JS}(q_i \mid\mid p_i)_{\mathrm{Ber}}.$$

Note that unlike the KL-divergence, the JS-divergence is symmetric in p and q, i.e. $\mathrm{JS}(p \mid\mid q) = \mathrm{JS}(q \mid\mid p)$.

3.3. Experiments and results

We use the Salient360! toolbox to calculate the metrics and included the implementation for the new proposed ones. Table 1 represents the models' scores on distinct datasets, including 2D models, 360° specialized models and the four baselines (illustrated in Figure 4). The ranking of the models is highly inconsistent across metrics, and no model is optimal for all metrics. Furthermore, the constant predictors achieve comparable scores with deep learning models, except on the $\mathrm{KL}_{\mathrm{Ber}}$ and $\mathrm{JS}_{\mathrm{Ber}}$. This supports our main claim: by assuming saliency maps as multivariate Bernoulli distributions in a principled way from model densities, deep learning models rank much better than the constant predictors.

Figure 3 shows the Spearman rank correlation across metrics. The goal is to cluster the metrics in terms of properties so those that favor/penalize similar behavior form a single cluster. All models from Table 1 are sorted according to each metric, then the Spearman rank correlation is measured between each pair of metrics to construct the covariance matrix in Figure 3. The NSS and CC represent the highest correlation (**0.96**), this can be explained by the fact that NSS is the discrete version of CC specifically designed for saliency. The pairwise correlation between AUC-J, NSS, CC, and SIM range between **[0.82,0.96]** meaning they capture the same axis of information and represent the same cluster.

KLD and JSD are sensitive to regularization and false negatives; this is beneficial if missing ground truth fixations should be severely penalized. Thus, pairwise correlation between KLD and the cluster ranges between 0.71 and 0.80. However, we demonstrated that just averaging the saliency maps on the training set can fool the metrics. For KLD, this can be explained as capturing the high energy and compact saliency region over the average scenes conditioned by the average fixation statistic; regularization to the unit sum will force the density to have one dominant mode and many smaller modes, thus, the regularization constant ψ in the metric computation is of the same value range as the majority

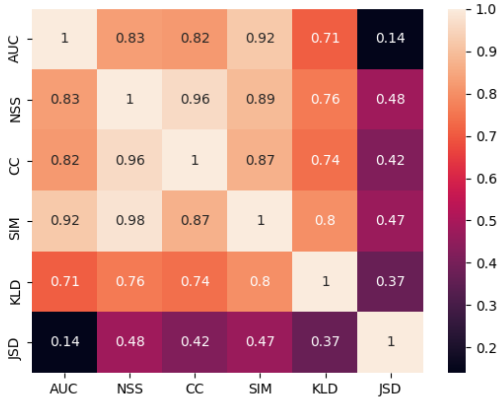

Figure 3. We sort the saliency models listed in Table 1 individually by each metric, and compute the Spearman rank correlation between every pair of metrics. The first 5 metrics listed are highly correlated. JSD is most uncorrelated with other metrics, due to their high sensitivity to zero-valued predictions at fixated locations.

Table 1. Comparative performance study on: Salient360! and Sitzman datasets. Baseline 01: the constant predictor from Salient360! dataset, Baseline 02: the constant predictor from Sitzman dataset, Baseline 03 is the Equator bias model. Baseline 04 is the random chance model.

Model		Salient360!								Sitzman							
		AUC-J ↑	NSS ↑	CC ↑	SIM ↑	KLD ↓	JSD ↓	KLD_B ↓	JSD_B ↓	AUC-J ↑	NSS ↑	CC ↑	SIM ↑	KLD ↓	KLD ↓	KLD_B ↓	JSD_B ↓
Human		0.788	2.09	1.0	1.0	0.0	0.0	0.0	0.0	0.985	3.421	1.0	1.0	0.0	0.0	0.0	0.0
2D models	UNISAL [4]	0.704	1.431	0.395	0.439	2.521	0.243	1.799	2.494	0.777	3.118	0.418	0.378	6.144	0.312	0.713	1.455
	SalGAN [18]	0.690	1.335	0.287	0.439	1.723	0.242	1.902	2.396	0.654	2.505	0.109	0.190	9.717	0.470	1.510	2.320
360° models	ATSAL [3]	0.777	1.658	0.642	0.639	0.761	0.111	2.916	4.501	0.802	3.216	0.461	0.360	6.584	0.321	1.240	4.166
	SalGAN360 [2]	0.774	1.621	0.613	0.617	0.881	0.125	1.143	1.295	0.830	3.136	0.428	0.387	6.176	0.302	1.184	1.617
Baselines	Baseline 01	0.768	1.635	0.593	0.603	1.022	0.132	2.327	2.920	0.777	2.978	0.340	0.295	7.827	0.375	1.729	5.196
	Baseline 02	0.725	1.530	0.491	0.550	1.744	0.170	2.414	1.978	0.735	2.948	0.321	0.301	7.939	0.373	1.5583	2.535
	Baseline 03	0.761	1.605	0.572	0.545	3.620	0.196	6.158	7.123	0.766	2.893	0.295	0.321	9.431	0.369	2.239	9.434
	Baseline 04	0.640	1.078	0.001	0.274	7.945	0.405	6.226	17.14	0.637	2.301	0.000	0.158	11.180	0.509	7.262	20.37

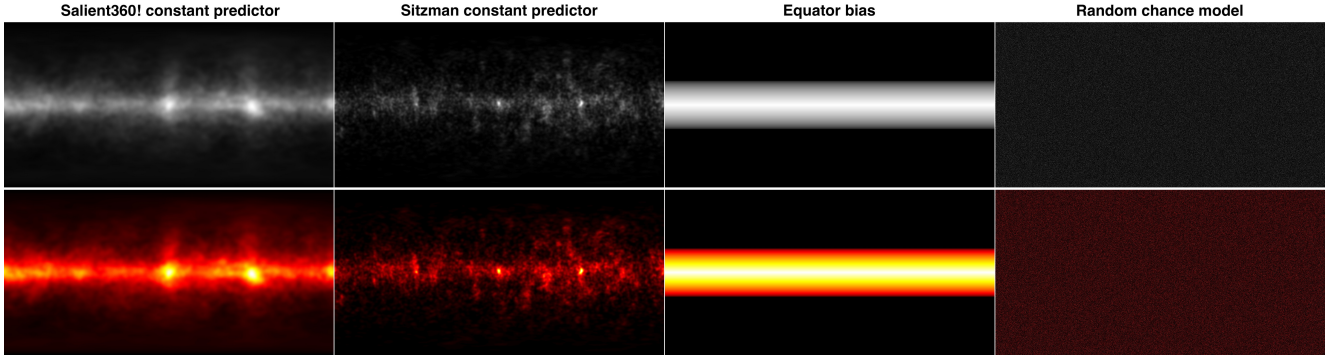

Figure 4. The Equi-rectangular saliency maps visualization of the four baselines. The upper row represent the density map, whereas the bottom row applies a color map to the saliency map for more perceptual figures.

saliency map values. The KLD score is influenced by the distance between the most dominant modes. This may interpret the good scores of the constant predictors. Although probabilistic, the Bernoulli assumption differs how JS_{Ber} and KL_{Ber} rank models because it modifies how saliency maps are defined: they measure the predictions locally pixel-wise. Despite the fact the pixels aren't fully independent, the Bernoulli assumption is less harmful and permits the metrics to fairly score all baselines. Furthermore, the pairwise correlation between JS_{Ber} and other metrics is low (**[0.14,0.48]**).

Figure 5 illustrates the metrics response per pixel-wise when comparing the ground truth saliency map to the models' saliency map. The red color indicates a worse score. In essence it shows that the metric incorrectly accredited the model for incorrect predictions (false positives/negatives). Ideally, a black map means a perfect classification from the metric. For KLD, it is therefore fooled by the true positives of the constant predictors; other regions (top and bottom) are not penalized effectively, so the final score is mostly dominated by the intersection points. This may be due to the regularization constant ψ that is included in the metric computation, and regularizing the input to be a valid probability distribution. Both KL_{Ber} and JS_{Ber} appear more stable, and better classify most pixels. Their local computation ensures these metrics do not favor the highest mode of the density from other modes. Each pixel is equally evaluated independent from the others. The fact that it expects the saliency value to be easily interpretable – as the probability that a fixation is expected to occur in that specific pixel – means that no softmax over pixels is required. This allows more stable computation of the metrics as fewer very small values occur in the map.

4. Discussion

Benchmarking saliency models is an open research question due to inconsistent rankings when varying the metrics. We define saliency metrics to be performance measures that assess saliency maps against ground truth fixations, and subsequently saliency maps to be multivariate Bernoulli distributions. We have shown that simple baselines can incorrectly obtain high scores for classical metrics, highlighting the fragility of the current paradigm and potential fallacious interpretation derived from benchmarking 360° head/eye movements prediction algorithms. The same models obtained low scores when presented as multivariate Bernoulli distributions and measured by probabilistic metrics validating this assumption. However, whilst the work reported here has addressed the issue raised regarding simple baselines, we believe that the underlying research question is far from solved. Despite the importance of visual attention modelling for 360° scenes for many tasks, because of the ambiguity related to benchmarks/metric, algorithmic development has

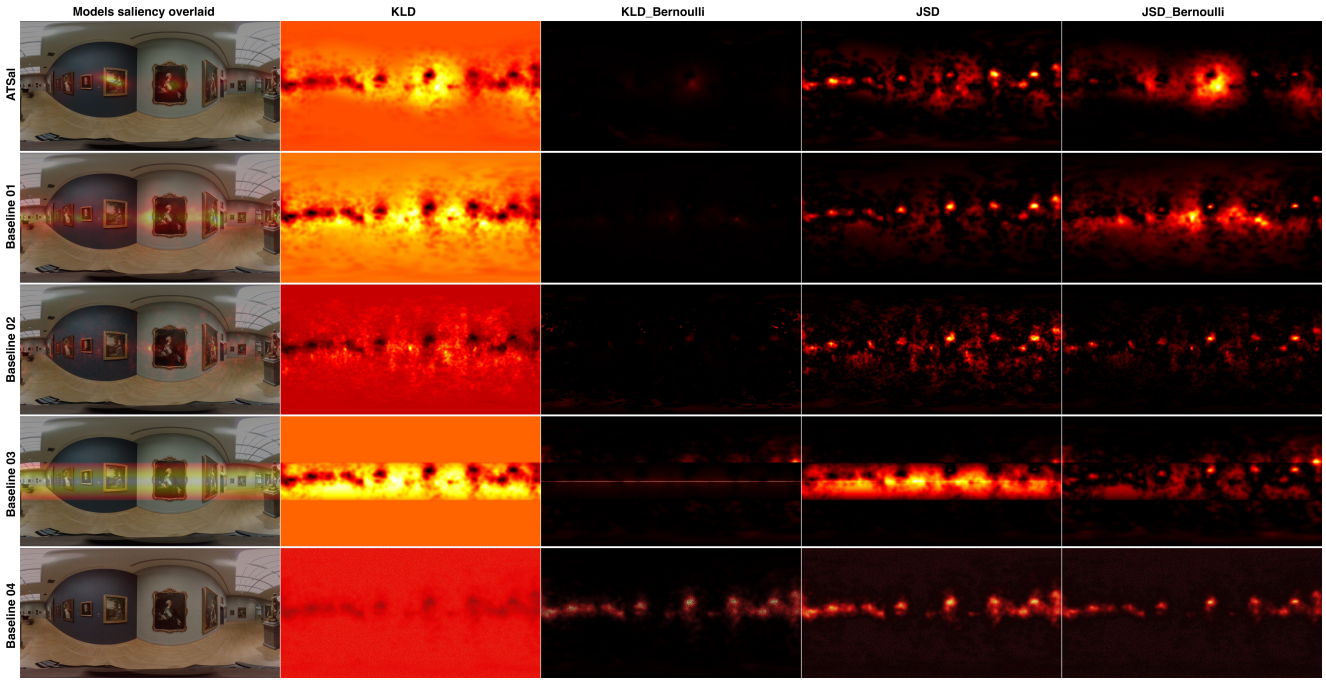

Figure 5. Visualizing metrics response can help us understand what behaviors of saliency models different evaluation metrics capture.

slowed. We call for reconsiderations on several aspects with benchmarking models.

Optimizing for the downstream task. Task-free saliency is the most common paradigm for most benchmarks. In practice, however, head/eye movement prediction is application dependant, such as in the case in visual quality assessment [5], or compression and transmission for 360° video/images [19]. Each task holds specific assumptions on the expected saliency map. For region based applications (e.g. compression and adaptive transmission), a location-based and local metric such as NSS would be more appropriate. However, if the task requires penalizing missing a fixation (e.g. object detection, surveillance, segmentation), metrics highly affected by false negatives (KLD and AUC) make more sense. This would require designing new task-specific datasets.

Decoupling the saliency model, map, and metric. Extending the work of [14] to 360° saliency with careful domain considerations would be interesting. Following the rationale of Bayesian decision theory: the saliency model is a posterior density over possible fixations and the saliency metric is a utility function. Based on the posterior density and the utility function, a saliency map is then chosen to maximize the expected utility. Thus, for each metric, a specific saliency map is derived from the model density. This has proved to be effective is getting a single winner on the 2D saliency benchmark MIT300 [12]. However, adapting the method to head/eye movements prediction requires rethinking the method developed in [13] in the context of omnidirectional datasets.

Considering the spherical geometry in the metric computation. The ground truth saliency maps are represented in the equi-rectangular format [7], which requires using helical sampling on the sphere. The metrics are calculated over the matching points at the polar positions. Thus, the scores are sensitive to the number of sampled points. Ideally, a measure should be invariant of the pre-processing steps. It is likely that representing the spherical specifications in the metric computation would be more effective.

5. Conclusion

Our work addresses the problem of benchmarking head and eye movements prediction models for 360° scenes. We identified a major failure mode for the classical metrics used in the current evaluation setups by using simple baselines. This means that deep learning based saliency models are no better than the average training saliency maps according to these metrics. Therefore, our analysis suggests defining saliency maps to be multivariate Bernoulli distribution as an alternative formulation. This is also supported by our understanding of how humans explore omnidirectional scenes. In practice, this means deriving new formulas for the probabilistic metrics. In this way, the baselines are detected to be poorer quality maps, whilst specialized models scores better across these probabilistic metrics. We will propose these remarks and recommendations to the Salient360! benchmark team.

Acknowledgement

This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant number 18/CRT/6183.

References

[1] 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.

[2] 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.

[3] Yasser Dahou, Marouane Tliba, Kevin McGuinness, and Noel O'Connor. Atsal: An attention based architecture for saliency prediction in 360 videos. In Alberto Del Bimbo, Rita Cucchiara, Stan Sclaroff, Giovanni Maria Farinella, Tao Mei, Marco Bertini, Hugo Jair Escalante, and Roberto Vezzani, editors, *Pattern Recognition. ICPR International Workshops and Challenges*, pages 305–320, Cham, 2021. Springer International Publishing.

[4] Richard Droste, Jianbo Jiao, and J Alison Noble. Unified image and video saliency modeling. In *European Conference on Computer Vision*, pages 419–435. Springer, 2020.

[5] Huiyu Duan, Guangtao Zhai, Xiongkuo Min, Yucheng Zhu, Yi Fang, and Xiaokang Yang. Perceptual quality assessment of omnidirectional images. In *2018 IEEE international symposium on circuits and systems (ISCAS)*, pages 1–5. IEEE, 2018.

[6] Mohsen Emami and Lawrence L Hoberock. Selection of a best metric and evaluation of bottom-up visual saliency models. *Image and Vision Computing*, 31(10):796–808, 2013.

[7] 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.

[8] Jesús Gutiérrez, Erwan J David, Antoine Coutrot, Matthieu Perreira Da Silva, and Patrick Le Callet. Introducing un salient360! benchmark: A platform for evaluating visual attention models for 360 contents. In *2018 Tenth International Conference on Quality of Multimedia Experience (QoMEX)*, pages 1–3. IEEE, 2018.

[9] Saumya Jetley, Naila Murray, and Eleonora Vig. End-to-end saliency mapping via probability distribution prediction. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5753–5761, 2016.

[10] Timothée Jost, Nabil Ouerhani, Roman Von Wartburg, René Müri, and Heinz Hügli. Assessing the contribution of color in visual attention. *Computer Vision and Image Understanding*, 100(1-2):107–123, 2005.

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

[12] Tilke Judd, Frédo Durand, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. In *MIT Technical Report*, 2012.

[13] Matthias Kümmerer, Thomas SA Wallis, and Matthias Bethge. Information-theoretic model comparison unifies saliency metrics. *Proceedings of the National Academy of Sciences*, 112(52):16054–16059, 2015.

[14] Matthias Kummerer, Thomas SA Wallis, and Matthias Bethge. Saliency benchmarking made easy: Separating models, maps and metrics. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 770–787, 2018.

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

[16] Jian Li, Martin D Levine, Xiangjing An, Xin Xu, and Hangen He. Visual saliency based on scale-space analysis in the frequency domain. *IEEE transactions on pattern analysis and machine intelligence*, 35(4):996–1010, 2012.

[17] Jia Li, Changqun Xia, Yafei Song, Shu Fang, and Xiaowu Chen. A data-driven metric for comprehensive evaluation of saliency models. In *Proceedings of the IEEE international conference on computer vision*, pages 190–198, 2015.

[18] 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.

[19] Sohee Park, Minh Hoai, Arani Bhattacharya, and Samir R Das. Adaptive streaming of 360-degree videos with reinforcement learning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1839–1848, 2021.

[20] 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.

[21] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *Proceedings of the IEEE international conference on computer vision*, pages 1153–1160, 2013.

[22] Benjamin W Tatler, Roland J Baddeley, and Iain D Gilchrist. Visual correlates of fixation selection: Effects of scale and time. *Vision research*, 45(5):643–659, 2005.

[23] Jozef L Teugels. Some representations of the multivariate bernoulli and binomial distributions. *Journal of multivariate analysis*, 32(2):256–268, 1990.

[24] Niklas Wilming, Torsten Betz, Tim C Kietzmann, and Peter König. Measures and limits of models of fixation selection. *PloS one*, 6(9):e24038, 2011.

[25] Mai Xu, Chen Li, Shanyi Zhang, and Patrick Le Callet. State-of-the-art in 360 video/image processing: Perception, assessment and compression. *IEEE Journal of Selected Topics in Signal Processing*, 14(1):5–26, 2020.