DEEP LEARNING VS. TRADITIONAL ALGORITHMS FOR SALIENCY PREDICTION OF DISTORTED IMAGES

Xin Zhao[1]*, Hanhe Lin*[2]*, Pengfei Guo*[3]*, Dietmar Saupe*[2] *and Hantao Liu*[1]

[1]School of Computer Science and Informatics, Cardiff University, United Kingdom
[2]Department of Computer and Information Science, University of Konstanz, Germany
[3]School of Computational Science, Zhongkai University of Agriculture and Engineering, China

ABSTRACT

Saliency has been widely studied in relation to image quality assessment (IQA). The optimal use of saliency in IQA metrics, however, is nontrivial and largely depends on whether saliency can be accurately predicted for images containing various distortions. Although tremendous progress has been made in saliency modelling, very little is known about whether and to what extent state-of-the-art methods are beneficial for saliency prediction of distorted images. In this paper, we analyse the ability of deep learning versus traditional algorithms in predicting saliency, based on an IQA-aware saliency benchmark, the SIQ288 database. Building off the variations in model performance, we make recommendations for model selections for IQA applications.

Index Terms— Image quality assessment, saliency, eye-tracking, distortion, statistical analysis

1. INTRODUCTION

Visual attention is one of the features of the human visual system that could extract most meaningful information in a visual field [1]. Knowing where people look in images helps understand how humans assess image quality [2]. Saliency – the stimulus-driven, bottom-up selective visual attention mechanism – has been integrated into various image quality assessment (IQA) algorithms to improve their performance [3], [4]. However, determining the optimal use of saliency in IQA algorithms is inconspicuous and depends on whether saliency can be reliably predicted in the IQA context. Unfortunately, there is a paucity of literature on saliency prediction for images that comprise various distortions.

In a representative study of image quality assessment, a set of pristine images is systematically degraded with diverse types and different levels of distortions. Previous IQA-aware eye-tracking studies [5], [6] have disclosed that distortions contained in an image cause the shifts or redistribution of saliency from its original places. This means the saliency of a distorted image differs from that of its corresponding pristine image, and the degree of the difference depends on the type and/or level of distortion. Therefore, being able to predict saliency of distorted images, in particular, the variation of saliency induced by distortion, is of fundamental importance to advanced IQA algorithms. Literature shows there are many saliency models already, however, they have been designed for training without explicitly considering the impact of image distortions. It is unknown yet whether these models can cope with the distortions added to the pristine images. Thus, the usefulness of these saliency models for IQA is worth a thorough investigation.

Recent advances in saliency modelling have demonstrated the plausibility of using computational technologies to predict human eye fixations [7]. A saliency model generally outputs a topographic map that represents conspicuousness of scene locations, where some parts of a scene that appear to an observer to stand out relative to their neighbouring parts. The algorithms for saliency prediction can be classified into deep learning and traditional methods. Traditional methods are based on extracting low-level image properties, such as intensity, colour and texture features, and combining (using mathematical rules and principles) these image features into a single topographical saliency map [8]–[12]. On the other hand, deep learning methods use convolutional neural networks for predicting saliency maps, where a saliency model is learned through constructing a deep network architecture [13]–[16]. Fig. 1 depicts saliency maps generated by different saliency models for images of high, medium and low perceptual quality. In computer vision applications, especially for salient object detection, the deep learning methods have been found more accurate than the traditional methods. However, it remains concealed to what extent both types of methods can predict the IQA-aware saliency.

In this paper, we carry out an evaluation of state-of-the-art saliency models, including 5 deep learning models and 5 traditional models by using an IQA-aware saliency benchmark, i.e., the SIQ288 database. Building on the results of our analyses and cross-comparisons, we offer guidelines for choosing saliency models and approaches for IQA applications.

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 251654672 – TRR 161 (Project A05).

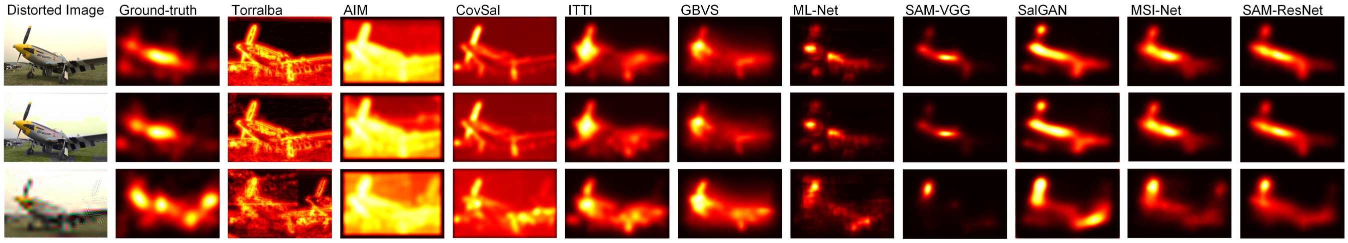


Fig. 1: Saliency maps (rendered by eye-tracking data or generated by different saliency models) for images of high (top row), medium (middle row), and low (bottom row) perceptual quality under distortion type "Fast Fading".

2. EXPERIMENTAL FRAMEWORK

2.1. SIQ288 saliency benchmark

This investigation uses the SIQ288 database [6] that consists of 288 images covering various artefacts and a diverse range of image qualities. A new experimental method was applied to reduce the inherent bias of saliency. Saliency maps were obtained via eye-tracking of 160 human observers. The images were selected from a benchmark image quality assessment database, the LIVE database [17]. The SIQ288 database contains 18 original images with each original image distorted into three perceptually distinctive levels (i.e., Low, Medium, and High distortion), and five different types of distortion (i.e., Fast Fading (**FF**), Gaussian Blur (**GBLUR**), JPEG Compression (**JPEG**), JPEG2000 Compression (**JP2K**), and White Noise (**WN**).

2.2. Description of visual saliency models

To perform a statistical analysis, we chose ten state-of-the-art saliency models from the MIT saliency benchmark [18]. They include five traditional models: Torralba [8], ITTI [9], GBVS [10], CovSal [11], and AIM [12]; and five deep-learning models: SAM-VGG [13], SAM-ResNet [13], ML-Net [14], SalGAN [15], MSI-Net [16].

The traditional models are based on different image features. Torralba contains both local and global-context features. ITTI combines multi-scale features of colours, intensity and orientations. GBVS is based on the graph theory. CovSal combines the local and global contrast. AIM computes saliency using Shannon's self-information measure of visual features. The deep learning models are based on different pre-trained neural network architectures. SAM-VGG (based on the VGG-16) and SAM-ResNet (based on the ResNet-50) use an Attentive Convolutional Long Short-Term Memory network to fine-tune the convolutional filters. ML-Net is a convolutional network that combines features extracted at different levels of the VGG16 network. SalGAN adopts convolutional encoder-decoder architecture and combines a unique adversarial loss function. MSI-Net is approached based on a convolutional neural network with encoder-decoder architecture that consists of multiple convolution layers at different dilution rates. Note, to make a fair comparative study, all models were implemented without re-calibration or retraining with the SIQ288 database.

2.3. Evaluation measures

Three commonly used evaluation metrics will be used to quantify the performance of saliency models. They are the value-based metric NSS, location-based metric AUC-Borji and distribution-based metric CC [19], [20].

Area under the curve (AUC Borji): This metric is one of the versions of the area under ROC (Receiver Operating Characteristic) curve measurement. The saliency map is used as a binary classifier to divide positive from negative samples at different threshold levels. AUC-Borji reduces the centre-bias to ensure a fair comparison of saliency models [19]. The range of AUC-Borji is from 0 to 1. The higher the value, the more accurate the saliency model predicts human eye fixations. Note, the AUC-Borji score of 0.5 indicates a random guess.

Normalised Scanpath Saliency (NSS): Normalised Scanpath Saliency metric measures the correspondences between the predicted saliency map (SM) and the ground truth fixations.

$$NSS(p) = \frac{SM(p) - \mu_{SM}}{\sigma_{SM}}, \tag{1}$$

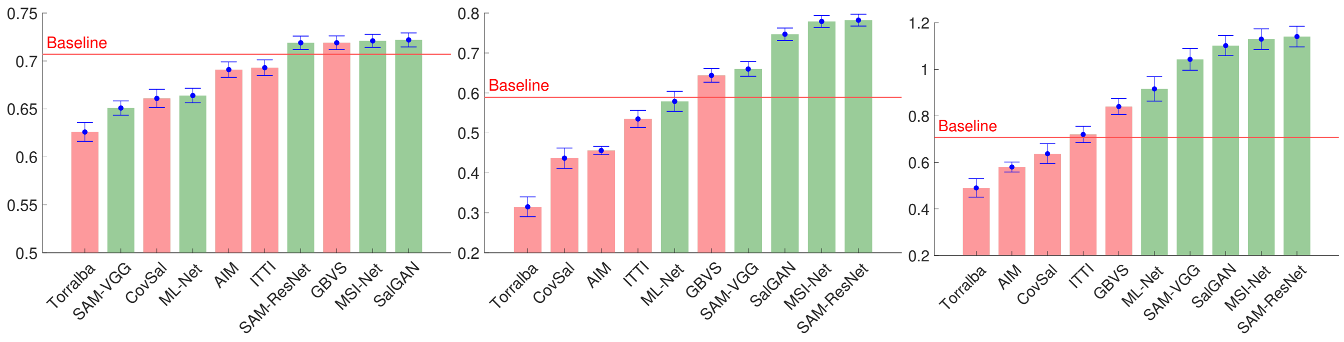
where p is the fixation location, σ_{SM} denotes the standard deviation of the SM, μ_{SM} denotes the average value.

The NSS score is the average of $NSS(p)$ for all fixations:

$$NSS = \frac{1}{N}\sum_{p=1}^{N} NSS(p), \tag{2}$$

where N is the total number of eye fixations.

When the score of NSS is bigger than 0, the greater the score, the higher the similarity between the predicted saliency map and the human fixations. If the NSS score is less than 0, it means the saliency map gives a very poor prediction.



(a) AUC-Borji

(b) CC

(c) NSS

Fig. 2: Performance of deep learning (in green bars) and traditional (in pink bars) saliency models measured by AUC-Borji (a), CC (b) and NSS (c) on the SIQ288 database. Error bars indicate a 95% confidence interval.

Pearson Linear Correlation Coefficient (CC): CC measures the linear correlation between the predicted saliency map PM and the ground truth saliency map SM:

$$CC(PM, SM) = \frac{\text{cov}(PM, SM)}{\sigma_{PM} \times \sigma_{SM}}, \tag{3}$$

where σ_{PM}, σ_{SM} denote the variance of PM and SM, and $\text{cov}(PM, SM)$ denotes the covariance of the two saliency maps. The range of CC value is between -1 and 1. When CC is close to 1 or -1, the two maps are highly correlated. The closer the CC is to 0, the less correlated are the two saliency maps. The value of zero indicates no correlation.

3. EXPERIMENTAL RESULTS

3.1. Overall performance

Based on the SIQ288 database, the ability of a saliency model to predict human gaze of distorted images is quantified by calculating AUC-Borji, NSS, and CC between the predicted saliency map and the ground truth eye-tracking data. Fig. 2 shows the rankings of saliency models' performance in terms of AUC-Borji, NSS, and CC, respectively. The baseline, as suggested in [18], indicates the performance of a "base" saliency model that is computed by stretching a symmetric Gaussian to fit the aspect ratio of a given image, under the assumption that the centre of the image is most salient. It can be seen from Fig. 2 that the cases above the baseline performance are dominated by the deep learning models, and there is only one traditional model, GBVS, performing above the baseline. The three deep learning models, i.e., SAM-ResNet, MSI-Net and SalGAN are consistently ranked higher than other models. In order to verify whether the difference in performance between traditional and deep learning models is statistically significant, hypothesis testing (by the independent samples t-test) is performed on the AUC-Borji, NSS, and CC data, using deep-learning/traditional as the independent variable. The results show that in all cases, the deep learning models are statistically significantly better than traditional models (i.e., AUC-Borji: $p < 0.05$, NSS: $p < 0.05$, and CC: $p < 0.05$).

3.2. Impact of distortion types

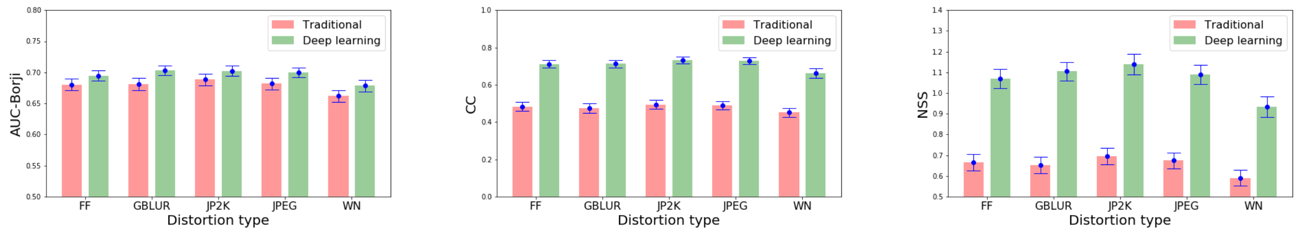
Now, we check the average performance of traditional versus deep learning models for different types of distortion contained in the SIQ288 database. Fig. 3 shows the results of performance comparisons based on AUC-Borji, NSS, and CC. It clearly indicates that deep learning models consistently outperform traditional models for all distortion types. We also performed an independent samples t-test for each comparison, and the results show that for each of the 15 cases (i.e., 5 types × 3 evaluation metrics) the difference is statistically significant ($p < 0.05$). Table 1 gives breakdown details of model performance. It can be also seen from the results that for the deep learning models, their performance on the WN distortion is relatively lower than other distortion types. The possible reason might be that how saliency is affected by distortion seems to be different for different distortion types.

3.3. Impact of distortion levels

Fig. 4 illustrates the average performance of traditional versus deep learning models for different levels of distortion (see also breakdown details in Table 2). The results of hypothesis testing (i.e., independent samples t-test) show that for each of the 9 cases (i.e., 3 levels × 3 evaluation metrics) the performance of the deep learning models is statistically significantly (i.e., $p < 0.05$) better than the traditional models. It is also worth noting here that although deep learning models are promising, they show relatively low performance in handling highly distorted images compared to images of low and medium levels of distortion. This might be due to these models having been trained with images without explicit distortions.

Table 1: Performance of individual saliency models measured by AUC-Borji, CC and NSS, for different distortion types.

	FF			GBLUR			JP2K			JPEG			WN		
	NSS ↑	CC ↑	AUC ↑	NSS ↑	CC ↑	AUC ↑	NSS ↑	CC ↑	AUC ↑	NSS ↑	CC ↑	AUC ↑	NSS ↑	CC ↑	AUC ↑
SAM-VGG	1.06	0.67	0.66	1.09	0.67	0.66	1.12	0.68	0.66	1.06	0.68	0.65	0.89	0.60	0.62
SAM-ResNet	**1.15**	**0.79**	0.72	**1.18**	**0.78**	0.72	**1.21**	0.80	**0.73**	**1.15**	**0.79**	**0.72**	1.02	**0.75**	**0.71**
ML-Net	0.93	0.59	0.67	0.97	0.60	0.67	0.99	0.61	0.67	0.97	0.62	0.68	0.71	0.47	0.63
SalGAN	1.08	0.74	**0.73**	1.11	0.74	**0.73**	1.16	0.75	0.72	1.12	0.76	**0.72**	**1.04**	**0.75**	**0.71**
MSI-Net	1.13	0.77	**0.73**	1.17	**0.78**	**0.73**	**1.21**	**0.81**	**0.73**	1.14	**0.79**	**0.72**	1.01	0.74	**0.71**
Torralba	0.49	0.32	0.62	0.45	0.29	0.62	0.55	0.35	0.64	0.52	0.33	0.63	0.44	0.29	0.61
ITTI	0.75	0.55	0.71	0.77	0.56	0.71	0.75	0.55	0.70	0.72	0.54	0.69	0.62	0.48	0.67
GBVS	0.86	0.65	**0.73**	0.87	0.65	**0.73**	0.85	0.64	0.72	0.84	0.64	**0.72**	0.81	0.65	**0.71**
CovSal	0.63	0.43	0.66	0.58	0.42	0.66	0.71	0.47	0.67	0.71	0.47	0.67	0.55	0.40	0.64
AIM	0.58	0.46	0.70	0.59	0.46	0.70	0.60	0.46	0.70	0.59	0.47	0.70	0.53	0.44	0.67

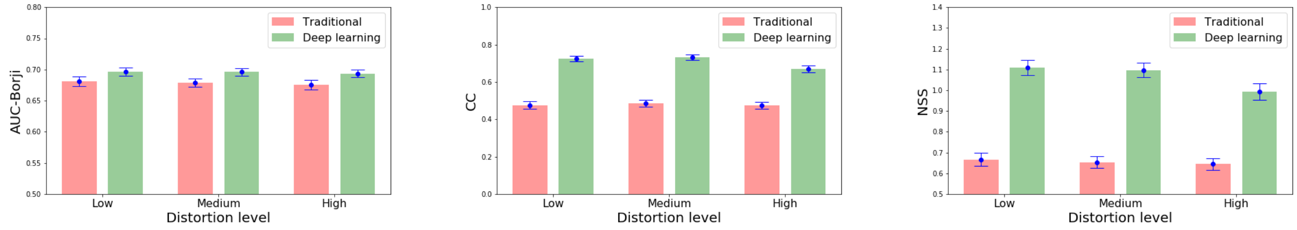


(a) AUC-Borji

(b) CC

(c) NSS

Fig. 3: Performance of deep learning (in green bars) and traditional (in pink bars) saliency models measured by AUC-Borji (a), CC (b), and NSS (c), for different distortion types. Error bars indicate a 95% confidence interval.



(a) AUC-Borji

(b) CC

(c) NSS

Fig. 4: Performance of deep learning (in green bars) and traditional (in pink bars) saliency models measured by AUC-Borji (a), CC (b), and NSS (c), for different distortion levels. Error bars indicate a 95% confidence interval.

Table 2: Performance of individual saliency models measured by AUC-Borji, CC and NSS, for different distortion levels. (i.e., Low, Medium and High distortion)

	AUC-Borji			CC			NSS		
	Low↑	Medium↑	High↑	Low↑	Medium↑	High↑	Low↑	Medium↑	High↑
SAM-VGG	0.66	0.65	0.64	0.67	0.68	0.63	1.08	1.07	0.98
SAM-ResNet	**0.72**	**0.72**	0.72	**0.79**	**0.80**	**0.75**	**1.17**	**1.16**	**1.09**
ML-Net	0.67	0.66	0.66	0.62	0.61	0.50	1.00	0.97	0.78
SalGAN	**0.72**	**0.72**	0.72	0.75	0.77	0.72	1.13	1.13	1.06
MSI-Net	**0.72**	**0.72**	0.72	**0.79**	**0.80**	0.74	**1.17**	**1.16**	1.06
Torralba	0.63	0.63	0.62	0.33	0.33	0.29	0.53	0.49	0.45
ITTI	0.69	0.69	0.70	0.52	0.54	0.55	0.71	0.72	0.74
GBVS	0.71	**0.72**	**0.73**	0.62	0.65	0.67	0.82	0.83	0.88
CovSal	0.67	0.66	0.65	0.45	0.44	0.41	0.69	0.64	0.58
AIM	0.70	0.69	0.69	0.46	0.47	0.44	0.59	0.58	0.57

4. CONCLUSION

In this paper, we conducted statistical analyses to evaluate the performance of deep learning versus traditional models for saliency prediction of distorted images. Obviously, deep learning models significantly outperform traditional models. In addition, we found that model performance tends to depend on the type and level of image distortion. Future work could focus on improving deep learning models for challenging cases, e.g., white noise distortion or highly distorted images.

5. REFERENCES

[1] L. Itti and A. Borji, “Computational models: Bottom-up and top-down aspects,” *Toet*, 2015.

[2] Z. Wang and A.C. Bovik, “Modern Image Quality Assessment,” *Synthesis Lectures on Image, Video, and Multimedia Processing*, vol. 2, no. 1, pp. 1–156, Jan 2006.

[3] H. Liu and I. Heynderickx, “Visual attention in objective image quality assessment: Based on eye-tracking data,” *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 21, no. 7, pp. 971–982, Jul 2011.

[4] W. Zhang, A. Borji, Z. Wang, P. L. Callet, and H. Liu, “The Application of Visual Saliency Models in Objective Image Quality Assessment: A Statistical Evaluation,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 6, pp. 1266–1278, Jun 2016.

[5] W. Zhang, Y. Tian, X. Zha, and H. Liu, “Benchmarking state-of-the-art visual saliency models for image quality assessment,” in *IEEE International Conference on Acoustics, Speech and Signal Processing*, May 2016, vol. 2016-May, pp. 1090–1094.

[6] W. Zhang and H. Liu, “Toward a Reliable Collection of Eye-Tracking Data for Image Quality Research: Challenges, Solutions, and Applications,” *IEEE Transactions on Image Processing*, vol. 26, no. 5, pp. 2424–2437, May 2017.

[7] A. Borji, D. N. Sihite, and L. Itti, “Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study,” *IEEE Transactions on Image Processing*, vol. 22, no. 1, pp. 55–69, Jan 2013.

[8] A. Torralba, A. Oliva, M.S. Castelhano, and J.M. Henderson, “Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search,” *Psychological Review*, vol. 113, no. 4, pp. 766–786, Oct 2006.

[9] D. Walther and C. Koch, “Modeling attention to salient proto-objects,” *Neural Networks*, vol. 19, no. 9, pp. 1395–1407, Nov 2006.

[10] J. Harel, C. Koch, and P. Perona, “Graph-based visual saliency,” in *Advances in Neural Information Processing Systems*, 2006, pp. 545–552.

[11] E. Erdem and A. Erdem, “Visual saliency estimation by nonlinearly integrating features using region covariances,” *Journal of Vision*, vol. 13, no. 4, 2013.

[12] N.D.B. Bruce and J.K. Tsotsos, “Saliency based on information maximization,” in *Advances in Neural Information Processing Systems*, 2005, pp. 155–162.

[13] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, “Predicting human eye fixations via an LSTM-Based saliency attentive model,” *IEEE Transactions on Image Processing*, vol. 27, no. 10, pp. 5142–5154, 2018.

[14] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, “A deep multi-level network for saliency prediction,” in *International Conference on Pattern Recognition*, Sep 2016, pp. 3488–3493.

[15] J. Pan, C.C Ferrer, K. McGuinness, N.E. O’Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto, “SalGAN: Visual Saliency Prediction with Generative Adversarial Networks,” Jan 2017.

[16] A. Kroner, M. Senden, K. Driessens, and R. Goebel, “Contextual Encoder-Decoder Network for Visual Saliency Prediction,” Feb 2019.

[17] H.R. Sheikh, Z. Wang, L. Cormack, and A.C. Bovik, “LIVE image quality assessment database release,” 2005.

[18] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, “Mit saliency benchmark,” http://saliency.mit.edu/.

[19] 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, 2019.

[20] 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 *IEEE International Conference on Computer Vision*, 2013, pp. 1153–1160.