Identifying Pitfalls in the Evaluation of Saliency Models for Videos

Zhengyan Dong*, Xinbo Wu*, Xin Zhao*, Fan Zhang† and Hantao Liu*
*School of Computer Science and Informatics, Cardiff University, Cardiff, CF24 4AX
†Department of Electrical and Electronic Engineering, University of Bristol, Bristol, BS8 1UB

***Abstract*—Saliency prediction has been extensively studied for natural images. In the area of video coding and video quality assessment, researchers attempt to integrate a saliency model to individual frames of a video sequence. In selecting best-performing saliency models for these applications, the evaluation only considers the average model performance over all frames of a video. This may mask the defects of a saliency model and consequently hinder further improvement of the model. In this paper, we present the identification of pitfalls in the evaluation of saliency models for videos. We demonstrate the importance of considering the video content classification and temporal effect. Building on the findings, we make recommendations for saliency model evaluation and selection for videos.**

***Keywords*—Saliency, eye-tracking, video, content classification, temporal effect**

I. Introduction

The past few decades have witnessed a significant growth in the use of digital videos in our daily lives. Videos are inevitably subject to distortions generated by compression and transmission. The distortions in video signals result in the reduction in video quality, which affects observers' visual experience and task performance [1]. To be able to control, monitor and improve quality of digital videos, a great deal of attention has been paid to the development of advanced algorithms for video compression and video quality assessment.

A current research trend in video compression and video quality assessment is to consider visual attention, which represents a powerful feature of the human visual system (HVS) [2], [3]. Visual attention mechanism enables the HVS to select the most relevant information from the visual scene. Simulating selective attention is highly beneficial for computational algorithms to distinguish between relevant and irrelevant visual signals and adaptively determine their parameters and processes. Many saliency models are available in the literature [4]. These models predict visual attention by generating a so-called saliency map, which represents conspicuousness of scene locations reflecting the relative importance of different image regions [5]. Saliency models are incorporated in video algorithms to produce saliency maps for individual frames (with or without temporal feature adaptation) [3], [6]. The frame-level saliency map can be used to weight the local algorithm output, for example, a distortion map calculated by a visual quality metric is weighted by a saliency map to generate a quality score for each frame, which is then averaged over all frames to generate a sequence-level quality score [7]. The effectiveness of these saliency-based video algorithms largely depends on the accuracy of the saliency model used [3].

Saliency model evaluation has been extensively studied for still images, where a saliency similarity score is computed between the predicted saliency and the ground truth [8]. However, the evaluation of saliency models for videos is less studied. The current practice is to sum up the frame-based evaluations and calculate a single score to represent the model accuracy for the entire video sequence. This evaluation regime neglects the temporal variations of saliency prediction accuracy and the impact of content classification on the overall performance of a saliency model. In this paper, based on the ground truth of eye-tracking data for videos [3], we perform statistical analyses to reveal the performance of state-of-the-art saliency models and identify pitfalls in the evaluation of saliency models for videos. Findings can help build reliable benchmark of saliency models for videos.

II. Problem Definition and Methodology

A. Eye-tracking data

The SVQ160 database [3] represents a reliable eye-tracking study, in which the data collection implemented rigorous control mechanisms to eliminate experimental biases. Note, the stimuli of the SVQ160 database contain both pristine and distorted videos. In this study, we only use the ten pristine videos with the aim to make the analyses more generally applicable, as the saliency of distorted videos are exclusively relevant for video compression and quality assessment. The reference videos include a diverse range of video content as shown in Fig.1. The videos are about ten seconds long and have a resolutions of 768×432 pixels. Eye movements of 20 observers were collected for each video. A frame-level saliency map (FSM) is generated from fixations over all subjects; and by each fixation giving rise to a Gaussian kernel that simulates the foveal vision (2° visual angle) of the HVS [3]. Fig.1 illustrates examples of the frame-level saliency maps.

B. Saliency models and performance measures

We selected a total of 14 state-of-the-art saliency models including seven traditional models and seven deep learning-based models. Table I lists these models and gives a brief description of each model.

A saliency model can be applied to each frame of a video sequence to generate a predicted frame-level saliency map (FSM). The prediction accuracy can be quantified by a similarity measure between the predicted FSM ($pred\_FSM$) and the ground truth FSM (gt_FSM). Amongst the popular saliency similarity measures [8], SIM (Similarity) and CC (Pearson linear correlation coefficient) have been proven to be the most

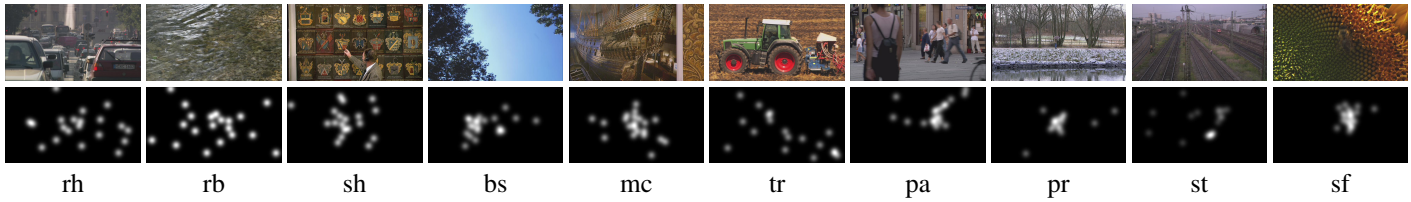

Fig. 1. SVQ160 database [3]: first row illustrates content (representative frames) of the original pristine videos, second row shows frame-level saliency maps.

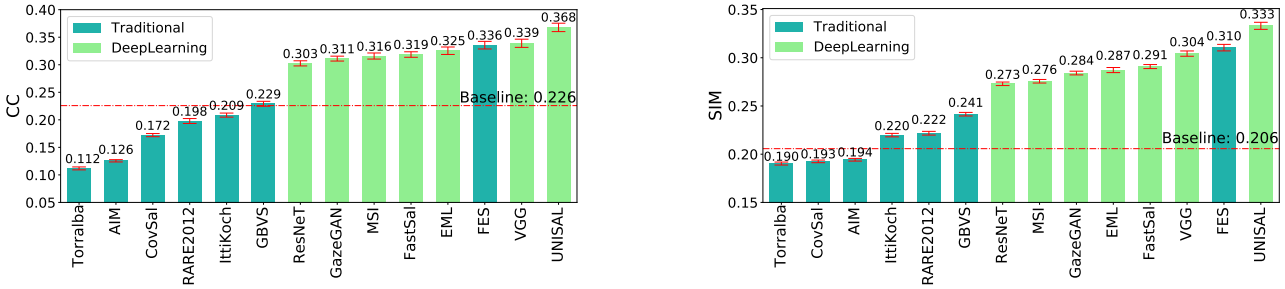

Fig. 2. Performance of state-of-the-art deep learning-based (in light green colour) and traditional (in dark green colour) saliency models (see details in Table I) measured by CC and SIM. Error bars indicate a 95% confidence interval. Baseline model is defined by "stretching a symmetric Gaussian to fit the aspect ratio of a given image, under the assumption that the center of the image is most salient [4]".

TABLE I. STATE-OF-THE-ART SALIENCY PREDICTION MODELS

Model name	Description (key words)
Traditional Saliency Models	
Rare2012 [9]	Multi-scale, rarity-based
AIM [10]	Information maximization
FES [11]	Sparse sampling, kernel density estimation
GBVS [12]	Graph-based
IttiKoch [13]	Colour, intensity, orientation
CovSal [14]	Region covariances
Torralba [15]	Local and global features
Deep Learning-based Saliency Models	
SAM-VGG & SAM-ResNet [16]	Long short-term memory (LSTM)
UNISAL [17]	Lightweight encoder-RNN-decoder
EML-Net [18]	Multilayer
FastSal [19]	MobielNet V2 backbone
MSI-Net [20]	Contextual encoder–decoder
GazeGan [21]	Generative adversarial network

appropriate perception-based measures for applications such as visual quality and compression [8], [22].

Similarity (SIM): SIM measures the similarity between the predicted and ground truth saliency maps when viewed as distributions $pred\_FSM_i$ and gt_FSM_i (equivalent to histogram intersection):

$$\mathrm{SIM}(\mathrm{pred\_FSM}, \mathrm{gt\_FSM}) = \sum_i \min\left(\mathrm{pred\_FSM}_i, \mathrm{gt\_FSM}_i\right) \quad (1)$$

where i represents the bin index of the histogram of a saliency map. The SIM's value range is between 0 and 1. The higher the SIM value is, the more accurate the saliency prediction is.

Pearson Linear Correlation Coefficient (CC): CC measures the linear correlation between the predicted saliency map $pred\_FSM$ and the ground truth saliency map gt_FSM:

$$\mathrm{CC}(\mathrm{pred\_FSM}, \mathrm{gt\_FSM}) = \frac{\mathrm{cov}(\mathrm{pred\_FSM}, \mathrm{gt\_FSM})}{\sigma_{\mathrm{pred\_FSM}} \times \sigma_{\mathrm{gt\_FSM}}} \quad (2)$$

where $\sigma_{\mathrm{pred\_FSM}}$, $\sigma_{\mathrm{gt\_FSM}}$ denote the standard deviation of pred_FSM and gt_FSM, respectively, and $\mathrm{cov}(\mathrm{pred\_FSM}, \mathrm{gt\_FSM})$ denotes the covariance of the two saliency maps. The range of CC is between -1 and 1. When CC is close to -1 or 1, the two maps are highly correlated meaning the saliency prediction is accurate. The closer the CC is to 0, the less correlated are the two saliency maps meaning the saliency prediction is less accurate.

C. Proposed saliency analysis methods

The goal of this paper is to show how saliency models behave for video applications. This can help identify appropriate metrics for saliency model evaluation and guide the selection of saliency models for videos. Existing saliency evaluation method is based on aggregating frame-level saliency over time and producing a single score to represent the prediction accuracy of the saliency model. This method ignores important properties of video saliency, including biases caused by visual content and temporal effect. To capture these properties for saliency evaluation, we consider the following methods:

Content-driven saliency dispersion (CSD): CSD [6] provides a quantitative measure for the degree of spatial saliency dispersion driven by visual content. Gaze is concentrated in fewer places in visual content with highly salient features than in content lacking salient features. Given a frame-level saliency map (FSM), CSD can be quantified by applying Shannon entropy to $p \times p$ non-overlapping blocks of the saliency map:

$$\mathrm{CSD}_i = H_{\sum}\left(\mathrm{FSM_i}\right) = \frac{1}{P_{\max}} \sum_{P=1}^{P_{\max}} \sum_{B=1}^{N_{\max}} H(B) \quad (3)$$

where H represents the entropy of a block B, $P_{\max}$ refers to the level of segmentation (i.e., $P_{\max} = 4$ was determined empirically in [6]). $N_{\max}$ is the $P_{\max}$ squared. The lower the CSD, the more concentrated the saliency is; otherwise, the higher the CSD, the more dispersed the saliency is in the spatial domain.

Temporal saliency prediction outliers (TSO): Temporal fluctuation of saliency prediction can significantly affect its application in video processing algorithms. A good saliency model should not only provide high time-aggregated accuracy, but also maintain its performance over time for a video sequence. To measure the temporal consistency of saliency prediction, TSO measures the ratio of outlier frames (N_{of}) to the total number of frames (N_{af}) of a video sequence. The outlier frames are the frames with a saliency prediction score (i.e., CC) below the threshold of $\text{CC}_{\text{mean}} - t \times \text{CC}_{\text{se}}$, where CC_{mean} and CC_{se} denote the mean and standard error of CC over all frames of the video (i.e, $t = 6$ was determined empirically in our experiment). TSO is defined as:

$$\text{TSO} = \frac{N_{\text{of}}}{N_{\text{af}}} \tag{4}$$

the lower the TSO value, the better the saliency prediction consistency is over time for the video.

III. Statistical analysis on saliency model behaviours

A. Time-aggregated model performance

First, we use the conventional method to evaluate saliency models for videos. For each video, the saliency performance measure (i.e., SIM or CC) is calculated for each frame, which is averaged over all frames to produce a performance value. Fig.2 shows the time-aggregated performance for the 14 saliency models as described in Section II. It can be seen that deep learning-based models outperform traditional models, except for the FES model. All deep learning-based models are better than the baseline model, which is defined by "stretching a symmetric Gaussian to fit the aspect ratio of a given image, under the assumption that the center of the image is most salient [4]". Now, we challenge this conventional saliency evaluation by identifying saliency model behaviours masked by this method. To reduce biases in our further investigation, we only select the models that are above the baseline in either of the rankings in Fig.2, including UNISAL, VGG, FES, EML, FastSal, MSI, GazeGAN, ResNet, GBVS. Note, for consistency CC is used as the saliency evaluation measure in the following analysis.

B. Impact of video content on saliency model performance

Hypothesis: We hypothesize that the impact of video content (VC) on the performance of saliency prediction models is statistically significant.

We first define the video content (VC) variable as a classification of the content-driven saliency dispersion (CSD) measure of equation (3). We calculate the sequence-level CSD by taking the average of frame-level CSD values. Fig.3(a) shows the saliency dispersion degree for all videos. Based on observed CSD values, we could classify the videos into two groups, i.e., VC_dispersed (including 'rh' to 'pa') and VC_compact (including 'pr', 'st' and 'sf'). In order to verify the content grouping is statistical meaningful, we perform hypothesis testing selecting CSD as the dependent variable and the categorical VC group as the independent variable. The Mann-Whitney U test is performed (due to evidence of non-normality as per the Shapiro-Wilk test) [23], and the results ($P < 0.05$) show that the CSD of group VC_dispersed is statistically significantly higher than that of group VC_compact as shown in Fig.3(b).

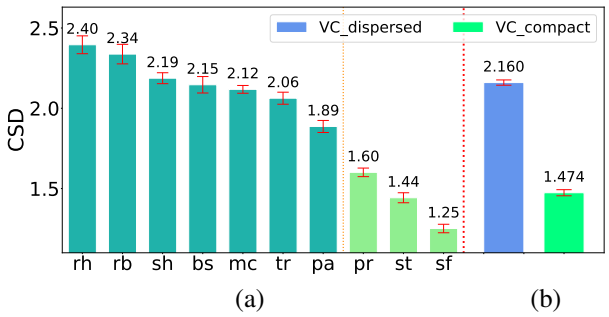

Fig. 3. Content-driven saliency dispersion (CSD) measure. (a) CSD values for individual videos. (b) CSD values for two distinctive video content categories, i.e., VC_dispersed versus VC_compact. Error bars indicate a 95% confidence interval.

Now, for the two distinctive video content classes (i.e., VC_dispersed and VC_compact), we analyse the impact of video content on the performance of saliency perdiction models. For each video, based on the frame-level CC of equation (2), we compute a sequence-level CC by averaging CC values over all frames. Therefore, 14 saliency models yield 14 sequence-level CC for each video. A hypothesis testing is conducted selecting sequence-level CC as the dependent variable, and the categorical VC group as the independent variable. The Mann-Whitney U test is performed (due to evidence of non-normality as per the Shapiro-Wilk test), and the results ($P < 0.05$) show that the model performance on group VC_dispersed is statistically significantly lower than that of group VC_compact, as shown in Fig.4(a). It can be seen that the top-performing saliency models tend to capture saliency of VC_compact videos but there is still room for improvement as the sequence-level $CC = 0.48$ remains inadequate as shown in Fig.4(a). However, these models fails in predicting the saliency of VC_dispersed videos (i.e., sequence-level $CC = 0.26$ indicates poor accuracy as shown in Fig.4(a)). The evidence implies that predicting saliency of complex scenes (as indicated by the VC_dispersed class, e.g., multiple objects) is more challenging than simple scenes (as indicated by the VC_compact class, e.g., a single object with dominant motion) for videos.

In addition, we analyse the individual saliency models in responding to VC_dispersed and VC_compact videos. The Mann-Whitney U test is performed (due to evidence of non-normality as per the Shapiro-Wilk test) on the sequence-level CC values produced by each model, and the results ($P < 0.05$) show that the model performance on VC_compact is statistically significantly higher than that of VC_dispersed for each model, as shown in Fig.4(b). It can be seen that for the VC_compact videos the model performance varies, e.g., UNISAL, VGG, FES and EML give a good prediction

TABLE II. TEMPORAL CONSISTENCY OF TOP-PERFORMING SALIENCY PREDICTION MODELS MEASURED BY TSO (TEMPORAL SALIENCY PREDICTION OUTLIERS). TEMPORAL-CONSISTENCY RANKINGS ARE COMPARED TO COMPARISON TO TIME-AGGREGATED RANKINGS.

Models	GAZEGAN	FASTSAL	ResNet	MSI	UNISAL	EML	FES	GBVS	VGG
TSO (temporal consistency)	0.365(1st)	0.371(2nd)	0.390(3rd)	0.380(4th)	0.387(5th)	0.388(6th)	0.393(7th)	0.398(8th)	0.410(9th)
CC (time-aggregated accuracy)	0.311(7th)	0.319(5th)	0.303(8th)	0.316(6th)	0.368(1st)	0.325(4th)	0.336(3rd)	0.229(9th)	0.339(2nd)

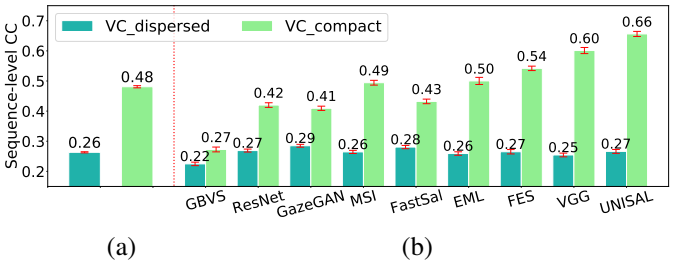

Fig. 4. Performance of top-performing saliency models measured by sequence-level CC. (a) Model performance for two distinctive video content categories, i.e., VC_dispersed versus VC_compact. (b) Individual model performance for two distinctive video content categories. Error bars indicate a 95% confidence interval.

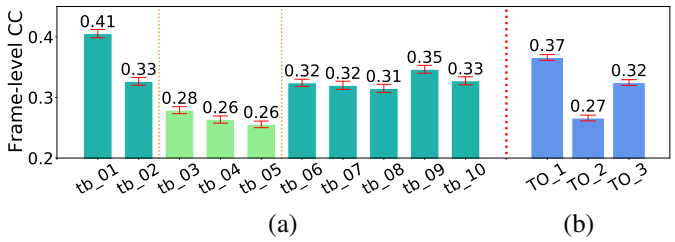

Fig. 5. Performance of saliency models over time. (a) Model performance in ten consecutive time blocks (i.e., tb_01 to tb_10). (b) Model performance in three semantic categories (beginning, middle and end) of time order (i.e., TO_1, TO_2, and TO_3). Error bars indicate a 95% confidence interval.

with CC larger than 0.5, and that all models consistently fail in predicting the saliency of the VC_dispersed videos. This suggests that without considering the impact of video content, bad model performance could be masked depending on the test database. In summary, the significant difference in model behaviours for different video content classes (i.e., simple or complex scenes) deserves more attention in both the evaluation of saliency models for videos and the construction of video eye-tracking databases, so that the biases could be account for in further research.

C. Impact of temporal context on saliency model performance

Hypothesis: We hypothesize that the impact of temporal context on the performance of saliency prediction models is statistically significant.

Little attention has been paid to the temporal variation of saliency model performance for videos. First, we measure the temporal consistency using TSO (temporal saliency prediction outliers) in equation (4). The TSO values of the nine top-performing saliency models are illustrated in Table II. In contrast to the time-aggregated performance values of Fig. 2(a), models rank high in the time-aggregated performance rankings do not necessarily give high performance consistency over time for videos, e.g., UNISAL ranks **1st** in the *time-aggregated rankings* but **5th** in the *temporal-consistency rankings*. The temporal consistency of saliency prediction is critical for applications such as video quality assessment and compression, where frame-based saliency weighting is often used [3].

Moreover, to analyse the model behaviours in the temporal context, we divide a video into ten consecutive blocks of time (i.e., each time block (TB) represents one second). For each time block, the mean of frame-level CC values (see equation (2)) over all videos (i..e, ten) and resulted from all (i.e., nine) saliency prediction models is computed. Fig.5(a) shows the model performance in time order (TO), indicating that the model prediction accuracy fluctuates over time for videos (e.g., a dip occurs around 3rd to 5th time blocks).

Based on the observation of Fig.5(a), we classify the time order (TO) into three semantic categories, including TO_1 (time blocks 1-2), TO_2 (time blocks 3-5), and TO_3 (time blocks 6-10). Hypothesis testing is conducted with frame-level CC as the dependent variable and the categorical TO as the independent variable. Pair-wise comparison is performed using the Mann-Whitney U test (due to evidence of non-normality according to the Shapiro-Wilk test). The results ($P < 0.05$) show that the difference in model performance between any two TO groups is statistically significant, as shown in Fig.5(b). The evidence indicates that the prediction accuracy of saliency models significantly deteriorates towards the middle section of a video sequence. A plausible reason could be that there is much uncertainty around the middle of viewing. In the beginning of the scene (i.e., TO_1), observers predominantly focus on salient regions; as time evolves (i.e., TO_2) observers' viewing behavior might change and gaze might be shifted from salient regions due to the tendency of exploring non-salient regions in the scene; after exploration observers would move gaze back to the salient regions during this space of time (i.e., TO_3). This speculation of observers' viewing behaviour could explain the extremely poor saliency model performance (i.e., $CC = 0.27$) for the middle section of a video sequence; meaning existing models cannot handle complex saliency shift. This temporal gaze behaviour poses challenges for accurately predicting saliency for videos. One way to address this problem is to include scene understanding components to saliency models, which is worth further investigation.

IV. CONCLUSION

In this paper, we formulate a new problem of saliency prediction – how to rigorously evaluate computational saliency models for videos. We found that video content has a significant impact on the performance of saliency models; and existing models fail in predicting saliency of videos of complex scenes. Also, the impact of temporal context on saliency model performance is significant; and existing models fail in capturing saliency in the middle section of a video. Findings can be used to facilitate the benchmark and selection of saliency models for video applications.

REFERENCES

[1] S. Winkler, *Digital video quality: vision models and metrics*. John Wiley & Sons, 2005.

[2] O. Le Meur, A. Ninassi, P. Le Callet, and D. Barba, "Overt visual attention for free-viewing and quality assessment tasks: Impact of the regions of interest on a video quality metric," *Signal Processing: Image Communication*, vol. 25, no. 7, pp. 547–558, 2010.

[3] W. Zhang and H. Liu, "Study of saliency in objective video quality assessment," *IEEE Transactions on Image Processing*, vol. 26, no. 3, pp. 1275–1288, 2017.

[4] Z. Bylinskii, T. Judd, A. Borji, L. Itti, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark. 2015," *URL: http://saliency. mit. edu/results_mit300. html*, vol. 12, p. 13, 2014.

[5] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[6] W. Zhang, R. R. Martin, and H. Liu, "A saliency dispersion measure for improving saliency-based image quality metrics," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 6, pp. 1462–1466, 2017.

[7] X. Feng, T. Liu, D. Yang, and Y. Wang, "Saliency inspired full-reference quality metrics for packet-loss-impaired video," *IEEE Transactions on Broadcasting*, vol. 57, no. 1, pp. 81–88, 2011.

[8] 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.

[9] N. Riche, M. Mancas, M. Duvinage, M. Mibulumukini, B. Gosselin, and T. Dutoit, "Rare2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis," *Signal Processing: Image Communication*, vol. 28, no. 6, pp. 642–658, 2013.

[10] N. Bruce and J. Tsotsos, "Attention based on information maximization," *Journal of Vision*, vol. 7, no. 9, pp. 950–950, 2007.

[11] H. Rezazadegan Tavakoli, E. Rahtu, and J. Heikkilä, "Fast and efficient saliency detection using sparse sampling and kernel density estimation," in *Scandinavian conference on image analysis*. Springer, 2011, pp. 666–675.

[12] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in neural information processing systems*, vol. 19, 2006.

[13] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[14] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *Journal of vision*, vol. 13, no. 4, pp. 11–11, 2013.

[15] 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, p. 766, 2006.

[16] 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.

[17] R. Droste, J. Jiao, and J. A. Noble, "Unified Image and Video Saliency Modeling," in *Proceedings of the 16th European Conference on Computer Vision (ECCV)*, 2020.

[18] S. Jia and N. D. Bruce, "Eml-net: An expandable multi-layer network for saliency prediction," *Image and Vision Computing*, vol. 95, p. 103887, 2020.

[19] F. Hu and K. McGuinness, "Fastsal: a computationally efficient network for visual saliency prediction," in *2020 25th International Conference on Pattern Recognition (ICPR)*. IEEE, 2021, pp. 9054–9061.

[20] A. Kroner, M. Senden, K. Driessens, and R. Goebel, "Contextual encoder–decoder network for visual saliency prediction," *Neural Networks*, vol. 129, pp. 261–270, 2020.

[21] Z. Che, A. Borji, G. Zhai, X. Min, G. Guo, and P. Le Callet, "How is gaze influenced by image transformations? dataset and model," *IEEE Transactions on Image Processing*, vol. 29, pp. 2287–2300, 2019.

[22] X. Yang, F. Li, and H. Liu, "A measurement for distortion induced saliency variation in natural images," *IEEE Transactions on Instrumentation and Measurement*, vol. 70, pp. 1–14, 2021.

[23] W. Zhang, A. Borji, Z. Wang, P. Le 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, 2015.