Optimizing Advertisement Placement Using Saliency Estimation in Filmmaking

Iakovos Raptis, Maria Tsourma, Anastasios Drosou, Dimitrios Tzovaras
Information Technologies Institute, *Centre for Research and Technology Hellas*, Thessaloniki, Greece
iakorapt@iti.gr, mtsourma@iti.gr, drosou@iti.gr, dimitrios.tzovaras@iti.gr

***Abstract*—As audiences spend an increasing amount of time consuming video content through online platforms, effective advertisement placement has become crucial for both engagement and revenue generation. While recommendation systems help match relevant advertisements to users, the optimal placement of film-related advertisements within visual content, remains underexplored. Traditionally, eye-tracking has been considered the most reliable method for gauging user attention, but it is both expensive and impractical in large-scale applications. This paper proposes a saliency-driven optimization approach for advertisement placement, leveraging saliency map estimations to predict areas of high visual attention. An automated solution that identifies optimal advertisement locations by focusing on the visual center of saliency maps is presented, eliminating the need for intrusive methods like eye-tracking. Additionally, a comparative analysis of two popular saliency estimation techniques is conducted, with both methods fine-tuned to enhance advertisement placement specifically in the context of film production. This approach aims to integrate advertisements seamlessly into visual media without disrupting viewer immersion. This approach will be used to insert film-related adverts in an audience engagement application.**

***Index Terms*—saliency, comparative study, ad placement**

I. Introduction

The effective placement of advertisements inside visual material poses a distinct problem and potential in the field of film production. Unlike static online advertisements, in-film advertisement (ad) placements have to blend in seamlessly without breaking the audience's immersion. There are now more opportunities for dynamic ad placement in films and videos because of the growth of digital content distribution and the move to online streaming platforms. Filmmakers and marketers may maximize the placement of advertisements in a way that both grabs viewers' attention and preserves the integrity of the viewing experience by utilizing cutting-edge technology like saliency estimate. Advertisements may be effortlessly included in the visual flow to generate more money without interfering with the viewing experience. Ads are placed in high-attention regions using strategies like saliency estimation, which makes them seem less obtrusive and more natural. This increases the efficacy of the advertisements and increases the likelihood that viewers will interact with them, which improves retention and may result in conversions. In addition, for distributors, ad placement provides flexibility, allowing for dynamic ads that can be tailored to different audiences, regions, or platforms. This increases the relevance of the ads while maintaining the integrity of the film's creative vision. By strategically placing ads, filmmakers can balance monetization with viewer satisfaction, ensuring that advertisements support rather than detract from the storytelling experience.

Traditional techniques of manual ad placement frequently fail to maximize the efficacy of ad campaigns, since they cannot easily gauge public interest. With the help of an AI-driven method, placement strategies may be more focused and successful, guaranteeing that advertisements appear in the most visually captivating parts of a movie, increasing their impact. AI-powered algorithms use a more sophisticated approach, analyzing massive quantities of data in real time to reach the appropriate audience with the right ad at the right moment. These models consider a variety of factors, including user behavior, demographics, geography, and even browsing history, to ensure that adverts are properly positioned to maximize engagement and conversions [6] [7].

By leveraging machine learning techniques, these algorithms can continuously learn and adapt to user preferences, delivering tailored advertisements that resonate with individual interests and needs [8]. This not only enhances the user experience by reducing irrelevant ad clutter but also increases the likelihood of driving conversions as users are more likely to engage with ads that are relevant to them. In an era where consumers are inundated with advertisements vying for their attention, personalization becomes a key differentiator in capturing and retaining audience interest.

Moreover, AI-based models offer scalability and efficiency in many domains [9]. These models can automate the process of ad selection and placement [3], freeing up valuable time and resources for website owners and advertisers to focus on other aspects of their marketing strategies. Additionally, AI algorithms can continually optimize ad placements based on real-time performance metrics, ensuring maximum ROI for advertisers while maximizing revenue potential for website owners. As the digital advertising landscape continues to evolve, embracing AI-based models for ad placement is no longer just a competitive advantage but a necessity for staying ahead in an increasingly dynamic and data-driven industry.

Apart from the recommendation of ads, one field related to the optimization and efficiency of the recommendation systems is related to the identification of effective locations within a web interface for the placement of content. Traditional approaches often rely on manual assessments or basic analytics

979-8-3315-0450-2/24/$31.00 ©2024 IEEE

to determine content placement, which can be time-consuming and subjective. AI-powered algorithms, however, offer a data-driven approach that analyzes numerous variables to pinpoint the optimal location for content dissemination.

This paper introduces an AI-based automated method for efficiently identifying optimal advertisement locations within web interfaces, utilizing an encoder-decoder neural network. The proposed system allows users to upload screenshots of their web interface, and the AI algorithm analyzes the visual layout to determine the most effective ad placement based on saliency estimation. By leveraging this technology, users can quickly identify prime ad locations without manual intervention, streamlining the process of ad positioning. Notably, this approach is adaptable to any web interface and does not require additional customization or specialized setup, making it widely applicable across different platforms.

In addition to presenting the implementation details of the saliency map estimation model, this paper also conducts a comparative analysis between the proposed method and other baseline approaches from existing literature. The comparison highlights the performance improvements and accuracy gains achieved through the use of the encoder-decoder neural network in saliency estimation. This evaluation not only demonstrates the effectiveness of the proposed model but also positions it as a competitive solution for enhancing advertisement placement in various digital contexts.

The rest of the paper is structured as follows: section II presents the related work available in the literature. Section III presents the network architecture proposed in this paper. Section IV presents the evaluation results of the proposed method and finally, section V presents the conclusions resulting from the implementation of the proposed architecture.

II. Related Work

There exist several studies that deal with the position and type of advertisement. Bo Ning et al. [1] examine the influence of structural and semantic factors of an advertisement on the visual attention and memory of said advertisement. The study concluded that an advertisement's structure and thematic consistency affect the visual attention it receives. In the realm of advertisement position, Saowwapak-adisak et al. [2] studied the impact of the position of a banner-type advertisement on brand awareness, visual fixation, and product knowledge. This study used an eye-tracking device to track the viewers' gaze and found that the banner's position affects the resulting product knowledge in some cases.

The research of Sulikowski et al. [3] explores elements related to a recommendation interface. They suggest a framework to assess the performance of a recommending interface, considering the unique characteristics and objectives of individual users. At the core of their approach is a deep neural network trained to predict the effectiveness of a specific recommendation, taking into account its position and intensity. The proposed framework, known as Performance Evaluation of a Recommending Interface (PERI), enables the automated adjustment of an optimal recommending interface based on user characteristics and goals. The experimental findings rely on data from the Gazepoint GP3 eye-tracker and synthetic data, used for pre-assessment training of the neural network.

There has been extensive research into saliency estimation models. Pan et al. [4] introduced SalGan, a deep convolutional neural network for saliency estimation that has been trained with an adversarial methodology. The initial phase of the network involves a generator model, and its weights are acquired through back-propagation using binary cross entropy (BCE) loss applied to downsampled versions of saliency maps. The outcome of this prediction undergoes processing by a discriminator network, which is trained for binary classification between the generatively produced saliency maps and the ground truth maps. Their experiments demonstrate that adversarial training, when coupled with a commonly used loss function such as BCE, enables achieving state-of-the-art performance across various metrics.

A neural model with a similar aim is Contextual Encoder-Decoder Network for Visual Saliency Prediction. Kroner et al. [5] introduced an approach utilizing a convolutional neural network pre-trained on a large-scale image classification task. The architecture adopted an encoder-decoder structure, featuring a module with multiple convolutional layers at varying dilation rates to capture multi-scale features simultaneously. Additionally, they integrated the obtained representations with global scene information to enhance the accuracy of visual saliency estimation. Their model demonstrates competitive and consistent performance across multiple evaluation metrics on two public saliency benchmarks, showcasing its effectiveness on five datasets and selected examples. In comparison to state-of-the-art methods, their network relies on a lightweight image classification backbone, making it a suitable choice for applications with limited computational resources, such as (virtual) robotic systems, aiming to estimate human fixations in complex natural scenes.

While related work does exist, it does not present a unified system or process that can process screenshots of websites and locate the optimal advertisement position based on visual saliency. The current paper proposes an automatic method for efficient ad location identification, and the proposed method is compared with the [4] and [5]. This comparison is being carried out in the domain of website screenshots using open and accessible datasets.

III. Methodology

The proposed solution is an automated method that calculates the optimal advertisement location (Figure 1). As an input, it accepts a screenshot of a website. This screenshot is being processed by the saliency estimation unit and produces a saliency map. A saliency map is a greyscale "heatmap" that showcases with white color the regions of interest, while the rest of the image is in black. These saliency maps correspond with their respective image. Using the resulting saliency map, the visual "center of weight" can be inferred using image moments of the first and zeroth order. An advertisement can be placed near or on top of this location, depending on the

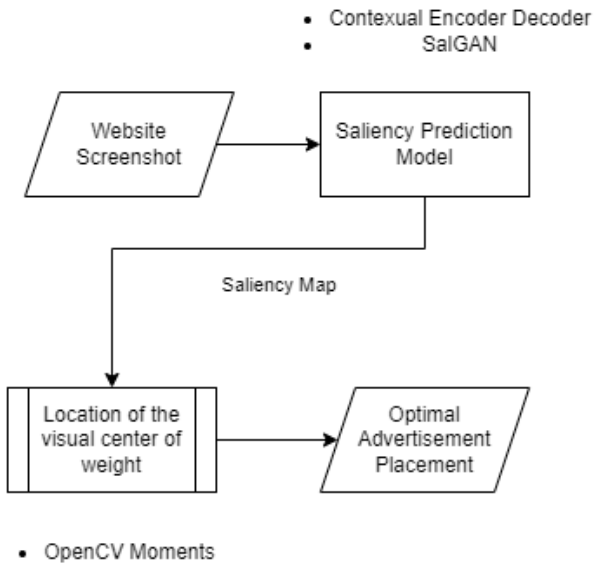

Fig. 1. The architecture of the proposed solution. Using a screenshot of a website, it outputs the optimal advertisement placement. The bullets signify the used methodology of their respective unit

will of the website developer, or the structure of the website. Moments can be calculated using the Equation 1.

$$M_{ij} = \sum_{x} \sum_{y} (x^i y^j I(x, y)) \tag{1}$$

In equation 1 Mij is the moment of (ij) order, x and y are the coordinates of a pixel and I(x,y) is the pixel's intensity on the x,y location.

Three different online, public datasets were used for the training of the two different deep learning models for saliency map estimation. The first one is the SALICON dataset [10], the second one is the Contrastive Websites, and the third one is the Gaze Mining dataset [13]. All three of them contain images and their respective saliency maps.

The initial training of both of the models happened using the SALICON dataset. The SALICON dataset consists of image pairs, with one being a photograph while the other is the saliency map. The saliency maps were created by gaze tracking of users that were presented with the photographs. It consists of 15000 image pairs, where the 10000 were used as the training set while the 5000 acted as a validation set.

Both the Contrastive Websites and the Gaze Mining datasets contain screenshots of internet websites and their respective saliency maps. These two datasets were combined and used to finetune the pre-trained models. The Gaze Mining dataset focuses on capturing gaze data to better understand visual attention patterns. This dataset collects data using eye-tracking technologies, capturing both fixation points (where a user's gaze lingers) and saccade points (rapid movements), allowing for detailed analysis of attention over time. In film production

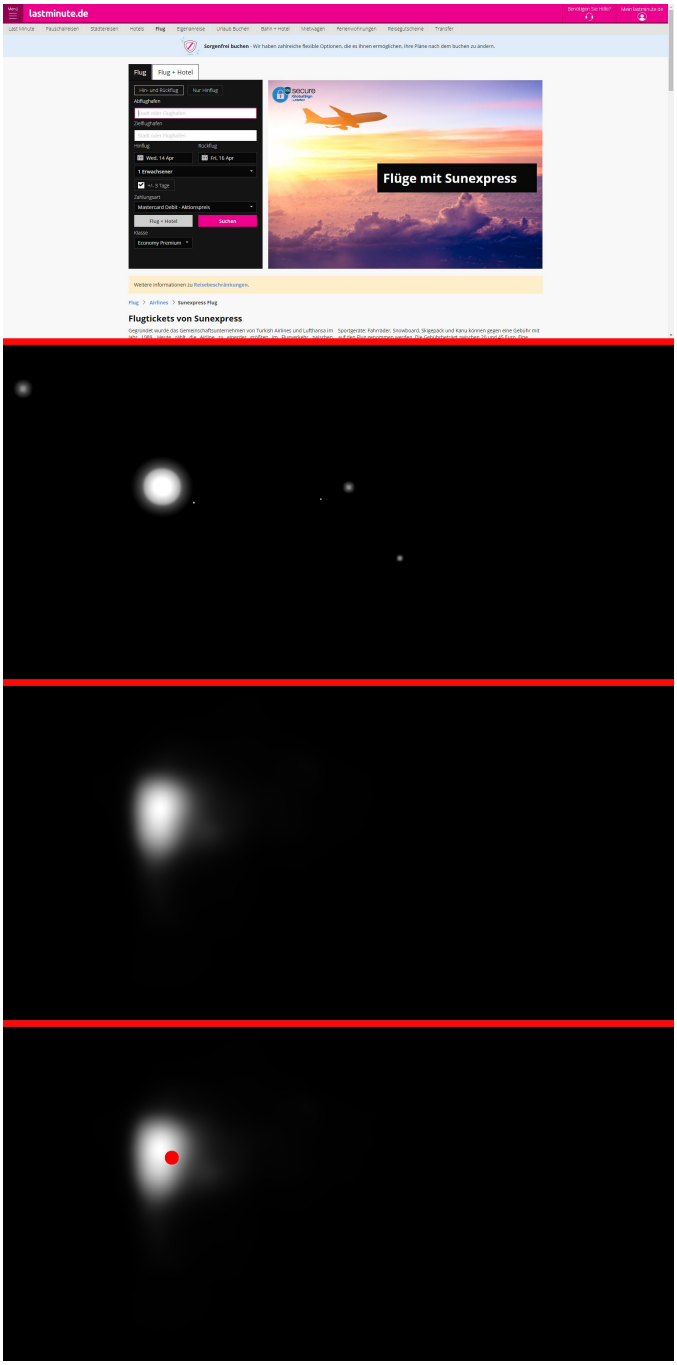

Fig. 2. Top image is a screenshot of a website. It acts as the input. The second image from the top is the ground-truth saliency map. The third image from the top is the estimated saliency map. The bottom image is the estimated saliency map with the visual center of weight highlighted with a red circle.

or web-based advertising, such a dataset can help optimize advertisement placement by understanding where viewers naturally focus their attention. The GazeFollow dataset is another relevant dataset. It consists of over 120,000 images annotated with gaze data from people looking at different scenes, making it suitable for analyzing visual attention in various contexts.

A. Contextual Encoder-Decoder Network for Visual Saliency Prediction

The proposed solution uses the Contextual Encoder-Decoder Network for Visual Saliency Prediction. It consists of an encoder and decoder architecture. Initially, it uses the pre-trained convolutional layers of VGG as a backbone. In the encoder part, there are three convolution layers with kernel sizes of 3x3 and dilation rates of 4,8 and 12, together with a 1x1 convolution layer. The output of layers 10, 14 and 18 is concatenated and used as input to an Atrous Spatial Pyramid Pooling module. After that, there is a convolution layer. The decoder consists of three convolutional and three upsampling layers to restore the original resolution of the image. In conclusion, the model consists of 24934209 trainable parameters.

B. Dataset

The training of the model was performed using the Salicon dataset [10]. The Salicon Dataset is a dataset produced by eye-tracking users while they view still photos. Salicon consists of pairs of images, with one image being the map that captures visual attention, and the second image being the image given to the user to measure visual attention using vision tracking techniques. Attention maps are black-and-white images, where any area that is white means it was noticed by users. The intensity of the white is proportional to the visual concentration of the users. This particular dataset is large enough to train the model satisfactorily for general use and to use this knowledge for more specific purposes with additional training [15].

IV. Evaluation

For the evaluation of the proposed model, multiple experiments were performed using the three different datasets. The initial training on the generic SALICON dataset, although offering competent results on generic photos and images, does not satisfy the specific use case of website advertising. Finetuning and transfer learning techniques were used to ensure that the knowledge gained in training with a large dataset is also useful in more specific cases. For the case of web screenshots, the two specific datasets were used to improve the overall accuracy.

A. Metrics

The metrics used are Mean Absolute Error (MAE), Mean Squared Error, and Structural Similarity Index. These metrics are widely used in saliency estimation and image similarity tasks.

a. Mean Absolute Error (MAE) is a widely used metric for evaluating saliency estimation models [11]. In the context of prediction, MAE measures the average absolute difference between the estimated saliency map and the ground truth saliency map at each pixel. This metric offers a simple and intuitive assessment of prediction accuracy, with smaller MAE values indicating better model performance. Since prediction mainly focuses on identifying the most visually significant regions in images, MAE effectively quantifies the model's ability to accurately approximate these regions. However, it does not take into account the spatial distribution or relative importance of different areas. Therefore, while MAE provides valuable information about prediction accuracy, it is often used together with other metrics to comprehensively evaluate the performance of significance prediction models.

$$MAE\left(P,GT\right) = \frac{1}{N}\sum_{i=1}^{N}\left|P_i - GT_i\right| \tag{2}$$

b. The mean squared error (MSE) [12] is another metric commonly used to evaluate predictive significance models. MSE measures the mean squared differences between the estimated saliency map and the ground truth saliency map at each pixel. While sharing similarities with the mean absolute error (MAE), MSE has the advantage of giving greater weight to larger forecast errors, making it more sensitive to outliers and extreme deviations. In the context of significance prediction, MSE can help identify areas where the model exhibits significant errors, emphasizing areas where the predicted significance values deviate greatly from the ground truth.

$$MSE\left(P,GT\right) = \frac{1}{N}\sum_{i=1}^{N}\left(P_i - GT_i\right)^2 \tag{3}$$

c. The Structural Similarity Index (SSIM) [14] is a valuable metric for evaluating significance prediction models. Unlike per-pixel metrics such as mean absolute error (MAE) or mean square error (MSE), SSIM takes into account not only differences in pixel values but also structural information, brightness, and contrast between estimated saliency maps and ground truth saliency maps. This makes SSIM more perceptual as it better aligns with human perception of image quality and similarity. When used to evaluate saliency estimation, a higher SSIM score indicates greater structural similarity between estimated saliency maps and ground truth maps, demonstrating the model's ability to capture not only pixel-level accuracy but also overall image structure. Thus, SSIM is a valuable addition to the evaluation toolbox for assessing the visual quality and structural coherence of significance predictions.

$$SSIM\left(P,GT\right) = \frac{\left(2\mu_P\mu_{GT} + c_1\right)\left(2\sigma_{P,GT} + c_2\right)}{\left(\mu_P^2 + \mu_{GT}^2 + c_1\right)\left(\sigma_P^2 + \sigma_{GT}^2 + c_2\right)} \tag{4}$$

B. Evaluation Results

The initial metrics come from early experiments showing the improvement in performance using different datasets. Initially, the proposed model was trained on the general Salicon dataset, then on the small but relevant Contrastive Websites dataset, and finally on the Combined Dataset that includes both the Contrastive Websites and the Gaze Mining datasets. The goal of this experiment is to understand and quantify the change in performance using different datasets. Subsequently, comparative tests were conducted using the SalGAN model.

Initially, the encoder-decoder model was trained on the general Salicon set. Then it was further trained on a more

targeted dataset (Contrastive Websites), and finally on the extended Combined dataset. Experiments were performed and the metrics of the results were recorded and are presented in Table 1 and Table 2.

TABLE I
METRICS OF EACH TRAINING SESSION USING THE CONTEXTUAL ENCODER- DECODER NETWORK

Metric	SALICON	CONTRASTIVE WEBSITES	COMBINED DATASET
MAE	0.1417	0.0580	**0.0398**
MSE	0.0553	0.0167	**0.0119**
SSIM	0.2900	0.4340	**0.5999**

TABLE II
PERCENTAGE IMPROVEMENT COMPARED TO THE INITIAL TRAINING ON SALICON (%) AFTER THE TWO FINETUNE TRAINING SESSIONS.

Metric	CONTRASTIVE WEBSITES	COMBINED DATASET
MAE	59.0185	71.9069
MSE	69.7024	78.4148
SSIM	49.6240	106.8475

Comparative tests were also conducted using a Generative Adversarial Network (GAN) type network. This model was trained for 240 epochs on the overall, generic Salicon dataset and then retrained using the Combined Dataset for 100 epochs. The results are shown in Table 3.

TABLE III
COMPARISON OF THE ENCODER-DECODER AND THE SALGAN MODELS IN THE DOMAIN OF WEBSITE SCREENSHOTS

Metric	Encoder Decoder	SalGAN
MAE	**0.040**	0.090
MSE	**0.012**	0.023
SSIM	**0.600**	0.353

V. CONCLUSION

This paper compared two distinct deep learning models and offered a method for determining the best place for advertisements on a website by utilising image processing and saliency estimation techniques. Although there are a lot of saliency estimate models that have already been trained, it is clear that further refined models on the necessary domain are needed. Our data unequivocally demonstrate that larger and more varied training sets produce superior outcomes. The comparison between the GAN model and the encoder-decoder model revealed that the encoder-decoder architecture is better appropriate for the current application. Across all computed metrics, the encoder-decoder model performs better than the others when it comes to assessing the saliency maps of webpages.

To improve the accuracy of our suggested solution even more, we intend to investigate different saliency map estimation techniques and increase the size of the training dataset in the future. The more varied and substantial the data set, the more the model will be able to generalize across various online interfaces. Furthermore, by modeling different scenarios and visual settings, new data augmentation techniques might increase the resilience of the model. Additionally, we intend to improve our strategy for mobile-only websites, modifying the solution to maximise ad placement in forms that are suitable to mobile devices, where screen real estate and user behaviour diverge greatly from desktop interfaces.

ACKNOWLEDGMENT

This project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No. 101095303 (project SCENE).

REFERENCES

[1] B. Ning, S. Luo, A. Wang, and M. Zhang, "Effects of banner ad type, web content type and theme consistency on banner blindness: an eye movement study," Cognitive Processing, vol. 24, no. 3, pp. 313–326, Aug. 2023, doi: 10.1007/s10339-023-01131-7.

[2] A. Saowwapak-adisak, J. Mongkolnavin, and P. Rattanawicha, "Impact Of Banner Ad Position, Congruence Of Banner Ad Content And Website Content, And Advertising Objective On Banner Ad Fixation, Brand Awareness, And Product Knowledge: An Empirical Study Of A Review Website In Thailand," 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:211837552

[3] P. Sulikowski and T. Zdziebko, "Deep Learning-Enhanced Framework for Performance Evaluation of a Recommending Interface with Varied Recommendation Position and Intensity Based on Eye-Tracking Equipment Data Processing," Electronics, vol. 9, no. 2, 2020, doi: 10.3390/electronics9020266.

[4] J. Pan et al., SalGAN: Visual Saliency Prediction with Generative Adversarial Networks. 2018.

[5] A. Kroner, M. Senden, K. Driessens, and R. Goebel, "Contextual encoder–decoder network for visual saliency prediction," Neural Networks, vol. 129, pp. 261–270, Sep. 2020, doi: 10.1016/j.neunet.2020.05.004.

[6] García-Sánchez, F., Colomo-Palacios, R., and Valencia-García, R. (2020). A social-semantic recommender system for advertisements. Information Processing and Management, 57(2), 102153.

[7] Verma, M., and Mishra, S. (2022, September). Recommendation Systems for Ad Creation: A View from the Trenches. In Proceedings of the 16th ACM Conference on Recommender Systems (pp. 525-527).

[8] Naumov, M., Mudigere, D., Shi, H. J. M., Huang, J., Sundaraman, N., Park, J., ... and Smelyanskiy, M. (2019). Deep learning recommendation model for personalization and recommendation systems. arXiv preprint arXiv:1906.00091.

[9] R. Y. Choi, A. S. Coyner, J. Kalpathy-Cramer, M. F. Chiang, and J. P. Campbell, Introduction to Machine Learning, Neural Networks, and Deep learning, Translational Vision Science and Technology, vol. 9, no. 2, pp. 14–14, 02 2020.

[10] M. Jiang, S. Huang, J. Duan, and Q. Zhao. SALICON: Saliency in Context. In Conference on Computer Vision and Pattern Recognition, Boston, 2015.

[11] A. Borji, H. R. Tavakoli, D. N. Sihite, and L. Itti. Analysis of Scores, Datasets, and Models in Visual Saliency Prediction. In ICCV, 2013.

[12] N. P. Hossein, Introduction to Probability, Statistics and Random Processes

[13] C. Zhang, D. Aspandi, and S. Staab, Predicting Eye Gaze Location on Websites. 2023.

[14] D. Brunet, E. R. Vrscay and Z. Wang, On the Mathematical Properties of the Structural Similarity Index, IEEE Transactions on Image Processing, vol. 21, no. 4, pp. 1488 - 1499, 2012

[15] Y. Luo, Y. Wong, M. S. Kankanhalli, and Q. Zhao, 'n-Reference Transfer Learning for Saliency Prediction', in Computer Vision – ECCV 2020, 2020, pp. 502–519.