CARDIFF UNIVERSITY
PRIFYSGOL CAERDYDD

ORCA – Online Research @ Cardiff

This is an Open Access document downloaded from ORCA, Cardiff University's institutional repository:https://orca.cardiff.ac.uk/id/eprint/165049/

This is the author's version of a work that was submitted to / accepted for publication.

Citation for final published version:

Lou, Jianxun, Wu, Xinbo, White, Richard, Wu, Yingying and Liu, Hantao 2024. Time-interval visual saliency prediction in mammogram reading. Presented at: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Seoul, Korea, 14-19 April 2024.

Publishers page:

Please note:
Changes made as a result of publishing processes such as copy-editing, formatting and page numbers may not be reflected in this version. For the definitive version of this publication, please refer to the published source. You are advised to consult the publisher's version if you wish to cite this paper.

This version is being made available in accordance with publisher policies. See http://orca.cf.ac.uk/policies.html for usage policies. Copyright and moral rights for publications made available in ORCA are retained by the copyright holders.

TIME-INTERVAL VISUAL SALIENCY PREDICTION IN MAMMOGRAM READING

Jianxun Lou[1], Xinbo Wu[1], Richard White[2], Yingying Wu[1], Hantao Liu[1]

[1]School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom
[2]Department of Radiology, University Hospital of Wales, Cardiff, United Kingdom

ABSTRACT

Radiologists' eye movements during medical image interpretation reflect their perceptual-cognitive behaviour and correlate with diagnostic decisions. Previous study has shown the significance of gaze behaviour of different time intervals for the decision-making process. Being able to automatically predict the visual attention of radiologists for different reading phases would enhance the reliability and explainability of artificial intelligence (AI) in diagnostic imaging. In this paper, we investigate the time-interval visual saliency in mammogram reading. We propose a novel visual saliency prediction model based on deep learning, which predicts a sequence of time-interval saliency maps for an input mammogram. Experimental results demonstrate the efficacy of the proposed time-interval saliency model.

Index Terms— Saliency, eye movements, deep learning, time-interval, mammogram

1. INTRODUCTION

Previous research has demonstrated that eye movements of radiologists when interpreting medical images provide insights into their perceptual-cognitive behaviour and decision-making process [1, 2]. Automatically predicating radiologists' visual attention during their diagnostic tasks benefits the advances of medical training [3] and artificial intelligence (AI)-aided diagnosis [4].

Visual attention representing the entire diagnostic process has been well studied; however, radiologists' visual attention for different time intervals during image interpretation is critical and less studied. Previous studies suggest that radiologists often fixate on abnormalities within the initial stages of image interpretation for example [5, 6]. The visual search patterns of radiologists commonly exhibit time-relevant features, due to their behaviour of adopting different search strategies during different reading stages and allocating varying levels of attention to abnormal regions over time [7]. Fig. 1 illustrates radiologists' visual attention for different time intervals when interpreting mammogram images, as well as the differences in visual attention patterns between these intervals. It can be seen that radiologists allocate their attention differently across different time intervals. Hence, the ability to automatically predict the visual attention of radiologists for different reading intervals is of importance for the development of AI in diagnostic imaging.

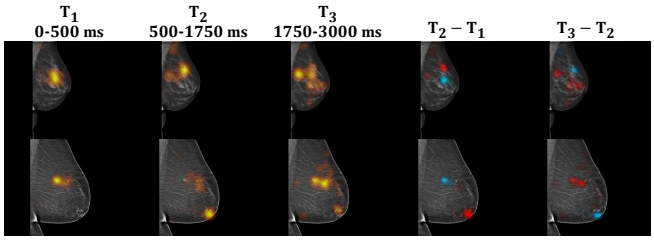

Fig. 1. Illustration of the radiologist's different visual attention allocations at different time intervals during reading mammograms. From left to right, the first three columns show saliency maps for T_1, T_2, and T_3 intervals. The following two columns represent attention difference for T_2-T_1 and T_3-T_2, where red regions signify emerging attention, and blue regions indicate decayed attention.

A range of saliency models have been developed and can effectively predict the visual saliency of images [8, 9, 10, 11, 12, 13]. Although the majority of these models are designed for natural images, some models are found to exhibit commendable performance in the domain of medical imaging [14]. It should be noted that these methods primarily focus on obtaining a holistic visual saliency map for the entire image reading process, hence neglecting the temporal information of the radiologist's visual behaviour during image interpretation. To capture temporal visual attention patterns, some studies have attempted to model scanpaths [15, 16], such as predicting the sequence of gaze locations observers fixate on an image over time. Despite this effort, the variability in individual fixation locations makes scanpath analysis and modelling notably complex. Furthermore, diverse sequences of content traversal in images might exhibit similar attention allocation in specific image regions. Recently, several studies have introduced modelling of temporal visual saliency on natural images [17, 18] to overcome these challenges. These methods maintain the robustness of population-level saliency while preserving the temporal characteristics of visual behaviour [17, 18]. While the methods for natural images have shown promise, medical images pose unique chal-

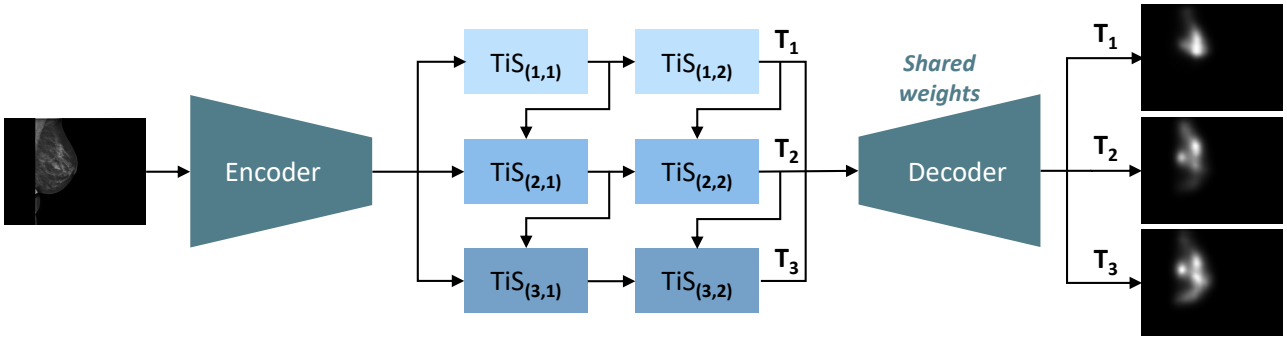

Fig. 2. Schematic overview of the architecture for the proposed method. A mammogram image is first encoded through an encoder, followed by three TiS modules. These TiS modules, from top to bottom, process the saliency features of time intervals 0-500 ms, 500-1750 ms, and 1750-3000 ms, respectively. The output features from each TiS module are then decoded by a shared-parameter decoder, producing the saliency map for each respective time interval.

lenges. It is beneficial to develop a method to predict temporal visual saliency of medical images to enhance the accuracy and clinical relevance of machine-generated saliency maps.

In this paper, we first explore the importance of predicting time-interval visual saliency during radiologists' mammogram examination, and then propose a novel model tailored for this task. This model leverages multiple time-interval saliency modules to achieve saliency prediction for mammograms. Experimental results demonstrate the effectiveness of the proposed model, and it outperforms the state-of-the-art visual saliency models for predicting time-interval saliency of mammograms.

2. PROPOSED METHOD

2.1. Time-interval saliency in mammogram reading

Our study is based on a large-scale mammography eye movement dataset comprising 196 mammogram images and eye-tracking data of 10 radiologists [19] scanning each image for 3000 milliseconds (ms). To explore the temporal saliency in mammograms, we divided the total reading time of 3000ms into three intervals: 0-500 ms (T_1), 500-1750 ms (T_2), and 1750-3000 ms (T_3). This interval division is based on the following reasons: First, for mammogram scanning, the eye movement behaviour of radiologists during the first 500 ms is significantly different from that in the period of 500-3000 ms [20]. Second, according to research [18], dividing time into equal intervals is effective for exploring the temporal aspects of visual saliency. The discrepancy in visual attention distribution between the different time intervals is assessed by calculating the Pearson's Correlation Coefficient (CC) and similarity (SIM) between the time-interval saliency maps [11]. The results in Table 1 show that the discrepancy between the time-interval saliency maps is evident (i.e. CC and SIM values are smaller than 0.7). Furthermore, statistical significant difference is found in CC and SIM for the comparisons of (T_1, T_2) versus (T_2, T_3), and (T_1, T_3) versus (T_2, T_3) (Mann–Whitney U tests, $p < 0.01$). These findings indicate profound differences in visual saliency across the T1, T2, and T3 time intervals, emphasising the importance of predicting time-interval visual saliency maps in mammograms.

Table 1. Discrepancy between saliency maps across different time intervals for mammogram images. T_1, T_2, and T_3 denote time intervals 0-500 ms, 500-1750 ms, and 1750-3000 ms, respectively.

	CC			SIM		
	T_1	T_2	T_3	T_1	T_2	T_3
T_1	1	0.5778	0.5706	1	0.4775	0.4634
T_2	0.5778	1	0.6391	0.4775	1	0.5529
T_3	0.5706	0.6391	1	0.4634	0.5529	1

2.2. Time-interval saliency prediction for mammograms

For temporal saliency prediction in mammograms, we propose a deep learning model, with its architecture illustrated in Fig. 2. In the proposed model, a mammogram image is first input into a static encoder for feature extraction; subsequently, the extracted features are fed into a series of *time-interval saliency* (TiS) modules to obtain visual saliency information for different time intervals; the output saliency maps are generated by a static decoder based on the saliency features of different time intervals.

Previous research has shown the pivotal role of encoders in predicting visual saliency for mammograms [14]. Following this approach, our model adopts the encoder as depicted in [14] to effectively extract mammogram image features, namely $\mathcal{F}_0$. In order to obtain the temporal visual saliency, $\mathcal{F}_0$ is processed by the TiS modules. Each TiS module leverages features from both previous iterations and the preceding time interval to derive saliency features for the current interval. Specifically, let $\text{TiS}_{(i,j)}$ represent the j-th layer of TiS

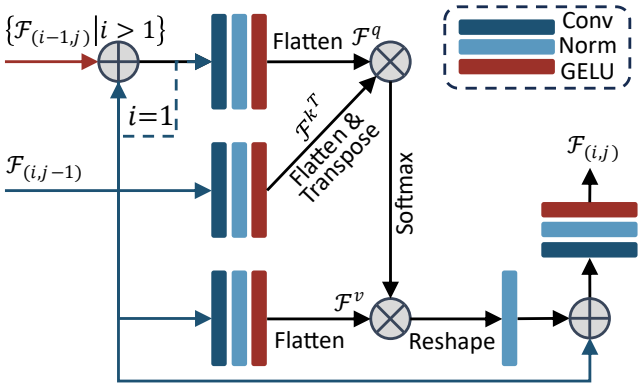

Fig. 3. Details of one time-interval saliency module ($\text{TiS}_{(i,j)}$).

applied to the i-th time interval, this process can be expressed as:

$$\mathcal{F}_{(i,j)} = \text{TiS}_{(i,j)}(\mathcal{F}_{(i-1,j)}, \mathcal{F}_{(i,j-1)}), \tag{1}$$

where $\mathcal{F}_{(i-1,j)} = \mathcal{F}_{(i,j)}$ when $i = 1$; $\mathcal{F}_{(i,j-1)} = \mathcal{F}_0$ when $j = 1$. For a single TiS module, the details are illustrated in Fig. 3. Let $g(\cdot)$ represent the sequence of convolution, normalisation, and GELU activation. In each TiS, for $i > 1$, we obtain $\mathcal{F}^q$ by flattening $g(\mathcal{F}_{(i-1,j)}+\mathcal{F}_{(i,j-1)})$, and $\mathcal{F}^k$ and $\mathcal{F}^v$ by flattening $g(\mathcal{F}_{(i,j-1)})$; for $i = 1$, all three are derived from flattening $g(\mathcal{F}_{(i,j-1)})$. Subsequently, the cross-interval attention can be computed using the following formula, yielding $\mathcal{F}^\star$ with the dimensions consistent with the input feature maps:

$$\mathcal{F}^\star = \text{rn}(\text{Softmax}(\mathcal{F}^q \times \mathcal{F}^{k^T}) \times \mathcal{F}^v), \tag{2}$$

where $\text{rn}(\cdot)$ represents reshape and normalisation. Finally, the final output of $\text{TiS}_{(i,j)}$ is:

$$\mathcal{F}_{(i,j)} = g(\mathcal{F}^\star + \mathcal{F}_{(i,j-1)}) \tag{3}$$

A static decoder is employed to reconstruct the outputs of these TiS modules into saliency maps for respective time intervals. The decoder is composed of a series of blocks, each containing a convolutional layer, normalisation, ReLU activation, and bilinear upsampling. These blocks reduce the channel dimensions of the saliency features for each time interval and adjust them to the spatial dimensions of the input mammogram image. In our proposed architecture, although separate saliency maps are generated for different time intervals, they utilise a common decoder with shared parameters.

3. EXPERIMENTAL RESULTS

3.1. Experimental settings

Previous studies [14] have indicated that network weights pre-trained on large-scale natural image saliency datasets are beneficial for visual saliency prediction of mammogram images. Accordingly, networks in our experiment were initialised with parameters pre-trained on SALICON [21], one of the most widely used and largest natural image saliency datasets. Thereafter, we employed 7-fold cross-validation on the eye-tracking mammogram dataset as described in section 2.1. More specifically, the dataset was partitioned into seven distinct subsets, each encompassing 28 images from 14 cases. In each run, a single subset was reserved for testing, another for validation, while the remaining five were amalgamated for training. Optimal models were derived using the early-stop strategy with patience of 5 epochs and were subsequently evaluated on the designated test set. The final results represent the mean performance across all seven runs. The loss functions used for each time interval are consistent and comprise a linear combination of CC, SIM, normalised scanpath saliency (NSS), and Kullback-Leibler divergence, as detailed in [14]. The optimisation process leverages the AdamW optimiser [22], commencing with an initial learning rate of 4×10^{-5}, which is multiplied by 0.1 every two epochs.

To evaluate the model performance in terms of the agreement between the predicted saliency maps and the ground truth, various evaluation metrics have been identified in the literature [23]. In this paper, four prevalent metrics, including CC, SIM, NSS, and area under ROC curve (AUC), are selected to render a holistic and impartial assessment for saliency prediction.

Table 2. Performance comparison between Baseline×3 and the variants of the proposed model. **N** represents the number of TiS used in each time interval.

		CC↑	SIM↑	NSS↑	AUC↑
T_1	Baseline×3	0.7971	0.6434	3.7957	0.9635
	N = 1	0.8109	0.6578	3.8583	0.9643
	N = 2	0.8152	0.6621	3.8885	0.9646
	N = 3	0.8162	0.6623	3.8943	0.9646
T_2	Baseline×3	0.8095	0.6799	2.9880	0.9468
	N = 1	0.8132	0.6844	3.0058	0.9469
	N = 2	0.8139	0.6849	3.0125	0.9469
	N = 3	0.8139	0.6850	3.0156	0.9469
T_3	Baseline×3	0.7811	0.6556	2.8219	0.9437
	N = 1	0.7827	0.6578	2.8294	0.9438
	N = 2	0.7831	0.6579	2.8295	0.9439
	N = 3	0.7818	0.6577	2.8270	0.9438

3.2. Effectiveness of time-interval saliency prediction

To evaluate the efficacy of our proposed method for predicting saliency across temporal intervals of mammograms, we defined a baseline model demoted as *Baseline×3*. Baseline×3 involves three instances of our model, but each having all TiS modules removed. Each of these instances was trained individually to predict the saliency map for a specific time interval. Let **N** represents the number of TiS modules of a single

Table 3. Performance comparison with state-of-the-art visual saliency models. The * symbol denotes static models, each comprising three model instances trained and tested individually on T_1, T_2, and T_3, while † denotes temporal saliency models. **Bold** indicates the best performance.

	T_1: 0-500 ms				T_2: 500-1750 ms				T_3: 1750-3000 ms			
	CC↑	SIM↑	NSS↑	AUC↑	CC↑	SIM↑	NSS↑	AUC↑	CC↑	SIM↑	NSS↑	AUC↑
DVA* [8]	0.7438	0.5737	3.4034	0.9584	0.7230	0.5879	2.5338	0.9336	0.7491	0.6218	2.5959	0.9386
UNISAL* [12]	0.7765	0.6251	3.6344	0.9605	0.7484	0.6355	2.6591	0.9376	0.7328	0.6166	2.5519	0.9364
EML-NET* [10]	0.7800	0.6167	3.6661	0.9614	0.7678	0.6443	2.8011	0.9412	0.7482	0.6226	2.6331	0.9376
MSI-Net* [13]	0.7902	0.6354	3.6515	0.9625	0.7941	0.6633	2.8368	0.9439	0.7650	0.6368	2.6750	0.9406
SAM-VGG* [9]	0.7917	0.6227	3.6218	0.9618	0.7739	0.6436	2.7063	0.9395	0.7235	0.5988	2.4525	0.9329
TranSalNet* [11]	0.7919	0.6379	3.7257	0.9619	0.7928	0.6662	2.9154	0.9444	0.7576	0.6371	2.7072	0.9400
TempSAL† [18]	0.8083	0.6555	3.8301	0.9636	0.7974	0.6710	2.8896	0.9449	0.7744	0.6489	2.7645	0.9421
Our Model†	**0.8152**	**0.6621**	**3.8885**	**0.9646**	**0.8139**	**0.6849**	**3.0125**	**0.9469**	**0.7831**	**0.6579**	**2.8295**	**0.9439**

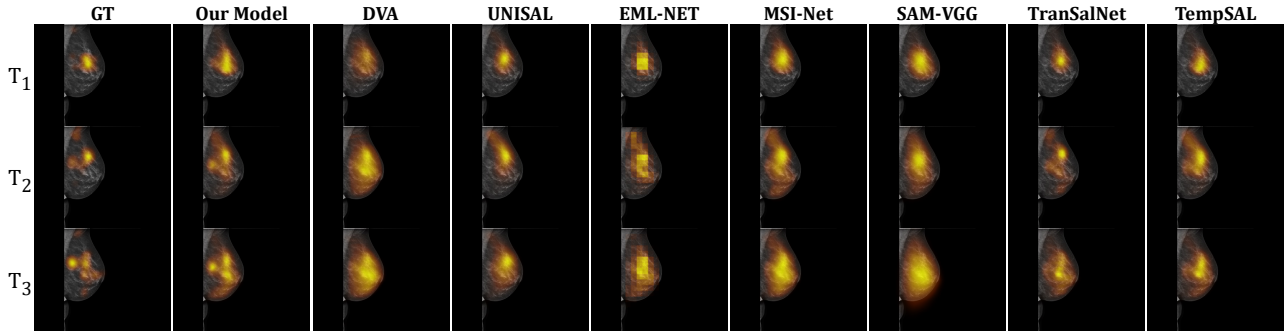

Fig. 4. Examples of time-interval saliency predictions. Rows, from top to bottom, represent time intervals: 0-500 ms, 500-1750 ms, and 1750-3000 ms. The leftmost column represents the Ground Truth (GT), the other columns show the predictions of state-of-the-art saliency models.

time interval. As can be seen from Table 2, by incorporating the TiS modules for time-interval saliency prediction, the proposed approach outperforms Baseline×3 across all metrics and time intervals. This indicates the efficacy of utilising the proposed method with TiS modules for time interval saliency prediction in mammogram images. Besides, the proposed model avoids the increased model parameters and computational consumption caused by training multiple model instances multiple times. Further comparison among the model variants with **N**=1, **N**=2, and **N**=3 reveals that overall the performance reaches a peak when **N**=2. To achieve optimal and consistent outcomes across the time intervals, we set **N**=2 in our proposed model.

3.3. Comparison with the state-of-the-art

To further substantiate the efficacy of the proposed method, we benchmark its performance against seven state-of-the-art visual saliency prediction models. Among these models, six are static models, including EML-NET [10], UNISAL [12], SAM-VGG [9], MSI-Net [13], DVA [8], and TranSalNet [11]; TempSAL [18] is a temporal saliency model. For a fair comparison, these models were first initialised by their pre-trained parameters and fine-tuned appropriately on the mammogram dataset with the same 7-fold cross-validation strategy as used for the proposed model. More specifically, for each static model, we instantiated three separate models, each trained independently to predict the saliency map for one specific time interval, similar to the approach taken with Baseline×3. For the temporal saliency model, the experimental approach applied was consistent with that of the proposed model. As demonstrated by the quantitative performance comparison in Table 3, the proposed model outperforms these state-of-the-art visual saliency models across all metrics. Additionally, Fig. 4 provides visualised results of the predictions. These results indicate the superior capability of the proposed method in predicting the time-interval saliency maps of mammogram images compared to other state-of-the-art models.

4. CONCLUSION

In this study, we have investigated the significance of the time-interval visual saliency in mammogram examinations conducted by radiologists. Then, we have developed a deep learning model that can accurately predict the time-interval saliency for mammograms. Experimental results have validated the effectiveness of our proposed model, and demonstrated its superior performance over existing state-of-the-art visual saliency prediction models.

5. REFERENCES

[1] Georgia Tourassi, et al., "Investigating the link between radiologists' gaze, diagnostic decision, and image content," *J. Amer. Med. Inform. Assoc.*, vol. 20, no. 6, pp. 1067–1075, 06 2013.

[2] Raymond Bertram, et al., "Eye movements of radiologists reflect expertise in CT study interpretation: A potential tool to measure resident development," *Radiol.*, vol. 281, no. 3, pp. 805–815, 2016.

[3] Damien Litchfield, et al., "Viewing another person's eye movements improves identification of pulmonary nodules in chest x-ray inspection," *J. Exp. Psychol. Appl.*, vol. 16, no. 3, pp. 251, 2010.

[4] Sheng Wang, et al., "Follow my eye: Using gaze to supervise computer-aided diagnosis," *IEEE Trans. Med. Imaging*, vol. 41, no. 7, pp. 1688–1698, 2022.

[5] Harold L Kundel, et al., "Holistic component of image perception in mammogram interpretation: gaze-tracking study," *Radiol.*, vol. 242, no. 2, pp. 396–402, 2007.

[6] Trafton Drew, et al., "Informatics in radiology: what can you see in a single glance and how might this guide visual search in medical images?," *Radiographics*, vol. 33, no. 1, pp. 263–274, 2013.

[7] Ziba Gandomkar, et al., "Visual search in breast imaging," *Br. J. Radiol.*, vol. 92, no. 1102, pp. 20190057, 2019.

[8] Wenguan Wang, et al., "Deep visual attention prediction," *IEEE Trans. Image Process.*, vol. 27, no. 5, pp. 2368–2378, 2018.

[9] Marcella Cornia, et al., "Predicting human eye fixations via an LSTM-based saliency attentive model," *IEEE Trans. Image Process.*, vol. 27, no. 10, pp. 5142–5154, 2018.

[10] Sen Jia, et al., "EML-NET: An expandable multi-layer network for saliency prediction," *Image Vis. Comput.*, vol. 95, pp. 103887, 2020.

[11] Jianxun Lou, et al., "TranSalNet: Towards perceptually relevant visual saliency prediction," *Neurocomputing*, vol. 494, pp. 455–467, 2022.

[12] Richard Droste, et al., "Unified image and video saliency modeling," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, pp. 419–435.

[13] Alexander Kroner, et al., "Contextual encoder–decoder network for visual saliency prediction," *Neural Netw.*, vol. 129, pp. 261–270, 2020.

[14] Jianxun Lou, et al., "Predicting radiologists' gaze with computational saliency models in mammogram reading," *IEEE Trans. Multimedia*, pp. 1–14, 2023.

[15] Olivier Le Meur, et al., "Methods for comparing scanpaths and saliency maps: strengths and weaknesses," *Behav. Res. Methods*, vol. 45, no. 1, pp. 251–266, 2013.

[16] Matthias Kümmerer, et al., "DeepGaze III: Modeling free-viewing human scanpaths with deep learning," *J. Vis.*, vol. 22, no. 5, pp. 7–7, 2022.

[17] Camilo Fosco, et al., "How much time do you have? modeling multi-duration saliency," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, 2020, pp. 4473–4482.

[18] Bahar Aydemir, et al., "TempSAL - uncovering temporal information for deep saliency prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recogn.*, June 2023, pp. 6461–6470.

[19] Lucie Lévêque, et al., "A statistical evaluation of eye-tracking data of screening mammography: Effects of expertise and experience on image reading," *Signal Process.: Image Commun.*, vol. 78, pp. 86–93, 2019.

[20] Jianxun Lou, et al., "Study of saccadic eye movements in diagnostic imaging," in *IEEE Int. Conf. Image Process.* IEEE, 2021, pp. 1474–1478.

[21] Ming Jiang, et al., "SALICON: Saliency in context," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1072–1080.

[22] Ilya Loshchilov, et al., "Decoupled weight decay regularization," *arXiv preprint arXiv:1711.05101*, 2017.

[23] Zoya Bylinskii, et al., "What do different evaluation metrics tell us about saliency models?," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, 2019.