PREDICTING RADIOLOGIST ATTENTION DURING MAMMOGRAM READING WITH DEEP AND SHALLOW HIGH-RESOLUTION ENCODING

Jianxun Lou[1], Hanhe Lin[2], David Marshall[1], Richard White[3], Young Yang[3], Susan Shelmerdine[4], Hantao Liu[1]

[1]School of Computer Science and Informatics, Cardiff University, Cardiff, United Kingdom
[2]National Subsea Centre, Robert Gordon University, Aberdeen, United Kingdom
[3]Department of Radiology, University Hospital of Wales, Cardiff, United Kingdom
[4]Department of Clinical Radiology, Great Ormond Street Hospital, London, United Kingdom

ABSTRACT

Radiologists' eye-movement during diagnostic image reading reflects their personal training and experience, which means that their diagnostic decisions are related to their perceptual processes. For training, monitoring, and performance evaluation of radiologists, it would be beneficial to be able to automatically predict the spatial distribution of the radiologist's visual attention on the diagnostic images. The measurement of visual saliency is a well-studied area that allows for prediction of a person's gaze attention. However, compared with the extensively studied natural image visual saliency (in free viewing tasks), the saliency for diagnostic images is less studied; there could be fundamental differences in eye-movement behaviours between these two domains. Most current saliency prediction models have been optimally developed for natural images, which could lead them to be less adept at predicting the visual attention of radiologists during the diagnosis. In this paper, we propose a method specifically for automatically capturing the visual attention of radiologists during mammogram reading. By adopting high-resolution image representations from both deep and shallow encoders, the proposed method avoids potential detail losses and achieves superior results on multiple evaluation metrics in a large mammogram eye-movement dataset.

***Index Terms*—** Eye movement, saliency, radiologist, mammogram, deep learning

1. INTRODUCTION

Humans have a tendency to focus their visual resources on the relevant visual information in a scene. Previous studies have demonstrated that the eye-movements of radiologists during diagnostic image reading can reflect their perception processes [1]. Being able to automatically predict the visual attention of radiologists would benefit their training, evaluation, and computer-aided diagnosis [2].

In image perception, visual saliency reflects the extent to which the content in a scene attracts visual attention. There have been several efforts to take advantage of the visual saliency of diagnostic images in various medical tasks and have obtained promising results. Banerjee et al. [3] utilised visual saliency to automatically segment brain tumors. Fan et al. [4] optimised the segmentation of dermoscopy images using visual saliency. Sran et al. [5] adopted visual saliency-based methods to detect experts' regions of interest when interpreting brain magnetic resonance images. These studies suggest that accurate prediction of the visual saliency of diagnostic images is of critical value to clinical practice.

Although various visual saliency prediction methods have achieved success in estimating visual attention, most of them have been designed for task-free viewing of natural images without specialised knowledge. Under these circumstances, humans tend not to only be attracted to low-level primitives such as contrast and luminance, but also bias their gaze on regions with familiar high-level semantics such as objects, faces, and animals [6]. In contrast, because radiologists are driven by their knowledge and visual search to make diagnoses on medical images, their eye-movement behaviours could be different from those for free viewing of natural images [7]. How to automatically predict the visual attention of radiologists during reading diagnostic images should be further explored.

Deep learning-based methods have gained success in the field of medical imaging [8]. However, for medical images consisting almost exclusively of low-level primitives, the informative details are highly likely to be ignored during the heavy downsampling that is commonly used in deep encoding images. Besides, previous study [9] has shown that shallow convolutional neural networks (CNNs) can encode radiological images' low-level primitives promisingly and mitigate the negative effects of data insufficiency on deep learning-based methods. Therefore, by minimising the potential detail losses, a combination of deep and shallow CNN encoders that provide higher-resolution image representations would boost the prediction of the radiologist's visual attention.

In this paper, we propose a novel deep learning-based method to automatically estimate the visual attention of radiologists during breast screening, which adopts high-resolution image representations from both deep and shallow encoders to minimize the potential detail losses. Also, we explore the impact of pre-training using a large-scale natural image saliency dataset on predicting the visual saliency of mammograms. The experimental results show that the proposed model outperforms the state-of-the-art on multiple

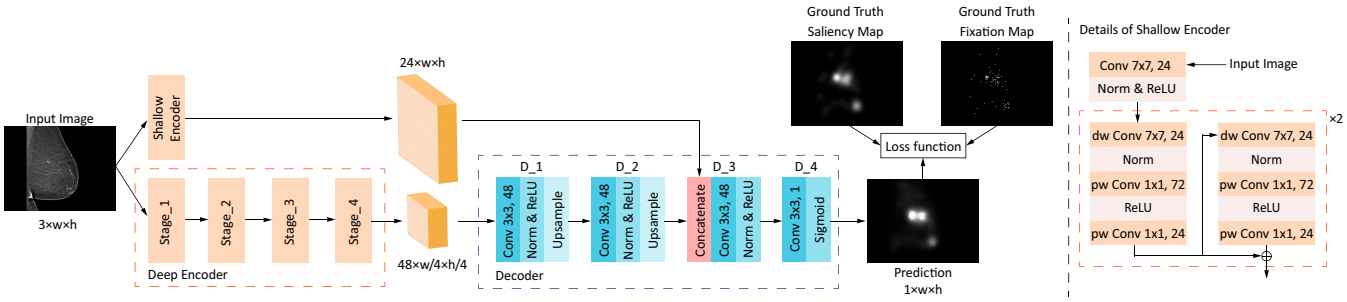

Fig. 1. An overview of the proposed method. Assume the spatial size of the inputs is $w \times h$. The input is first processed by the deep encoder and shallow encoders in parallel. These encoders provide high-resolution features with the spatial size of $\frac{w}{4} \times \frac{h}{4}$ and $w \times h$, respectively, which aims to minimize the potential detail losses. The structure of the shallow encoder is shown on the right. The final saliency map is generated by the decoder based on the features from both the deep and shallow encoders.

metrics in a benchmark mammogram eye-tracking dataset.

2. PROPOSED METHOD

2.1. Architecture

The proposed method is a deep learning-based model with an encoder-decoder architecture to predict radiologists' visual attention when interpreting diagnostic images. The architecture of the proposed model is shown in Fig. 1.

In the encoder phase, two parallel CNN-based encoder networks, including a deep encoder and a shallow encoder, are applied to extract image features. The deep encoder is based on *HRNet* [10] to provide high-resolution image features. Assume that the spatial size of the inputs is $w \times h$. Different from the encoder typically used in salency prediction that provides low-resolution image features (e.g., $\frac{w}{8} \times \frac{h}{8}$ [11, 12] or $\frac{w}{32} \times \frac{h}{32}$ [13, 14, 15] of the space size of the output in the final stage), this encoder provides image representations in $\frac{w}{4} \times \frac{h}{4}$ spatial size at the end of its final stage for the downstream decoder. The shallow encoder is used to extract high-resolution image features at a shallow depth, which does not employ any downsampling to obtain high-resolution representations of the input image (i.e., keeps the spatial size of $w \times h$ consistently). The architecture of the shallow encoder is shown on the right-hand side of Fig. 1. Because there is no downsampling in this encoder, a large size of convolution kernels (i.e., 7×7) is adopted to obtain a wider receptive field. The input images are first processed by a convolution with the kernel size of 7×7 and then passed into two residual blocks with the same architecture. Each residual block includes two identical sub-blocks linked by residual connections. For each sub-block, the feature maps are first processed by depthwise convolution with the kernel size of 7×7. After that, their dimensions are increased and then restored by two layers of convolution with the kernel size of 1×1. Previous studies have demonstrated that this kind of so-called "inverted residual" structure can enhance the representational capabilities of neural networks [16].

In the decoder phase, the outputs of the deep encoder are processed in series by two convolutions with linear upsampling to restore them to the initial spatial size. After that, the upsampled feature maps and the outputs from the shallow encoder are concatenated and sent to the subsequent networks to fuse together, which allows the model to utilize high-resolution image representations from the shallow encoder to support the saliency estimation.

2.2. Loss function

Using the saliency evaluation metrics to define the loss function has achieved notable success in saliency prediction [11, 15, 13, 12, 14]. Accordingly, we adopted a linear combination of four metrics as the loss function to train our model, including the normalized scanpath saliency (NSS), Kullback-Leibler divergence (KLD), linear correlation coefficient (CC), and similarity (SIM). Let $\mathbf{y}^s$, $\mathbf{y}^f$, and $\hat{\mathbf{y}}$ be the ground truth saliency map, fixation map, and predicted saliency map, and i indicates the ith pixel of $\mathbf{y}^s$ and $\hat{\mathbf{y}}$, our loss function is defined as:

$$\begin{aligned}\mathcal{L}(\mathbf{y}^s, \mathbf{y}^f, \hat{\mathbf{y}}) =& \lambda_1 \mathcal{L}_{\text{NSS}}(\mathbf{y}^f, \hat{\mathbf{y}}) + \lambda_2 \mathcal{L}_{\text{KLD}}(\mathbf{y}^s, \hat{\mathbf{y}}) \\ &+ \lambda_3 \mathcal{L}_{\text{CC}}(\mathbf{y}^s, \hat{\mathbf{y}}) + \lambda_4 \mathcal{L}_{\text{SIM}}(\mathbf{y}^s, \hat{\mathbf{y}}),\end{aligned} \tag{1}$$

where λ_1, λ_2, λ_3, and λ_4 are the weights of individual metrics, and

$$\mathcal{L}_{\text{NSS}}(\mathbf{y}^f, \hat{\mathbf{y}}) = \frac{1}{\sum_i \mathbf{y}_i^f} \sum_i \frac{\hat{\mathbf{y}}_i - \mu(\hat{\mathbf{y}})}{\sigma(\hat{\mathbf{y}})} \mathbf{y}_i^f, \tag{2}$$

where $\sigma(\cdot)$ and $\mu(\cdot)$ stand for standard deviation and mean respectively;

$$\mathcal{L}_{\text{KLD}}(\mathbf{y}^s, \hat{\mathbf{y}}) = \sum_i \mathbf{y}_i^s \log(\epsilon + \frac{\mathbf{y}_i^s}{\epsilon + \hat{\mathbf{y}}_i}), \tag{3}$$

where ϵ is a regularization constant and set to 2.2204×10^{-16};

$$\mathcal{L}_{\text{CC}}(\mathbf{y}^s, \hat{\mathbf{y}}) = \frac{\text{cov}(\mathbf{y}^s, \hat{\mathbf{y}})}{\sigma(\mathbf{y}^s)\sigma(\hat{\mathbf{y}})}, \tag{4}$$

where $\mathrm{cov}(\cdot)$ is the covariance and $\sigma(\cdot)$ is standard deviation;

$$\mathcal{L}_{\text{SIM}}(\mathbf{y}^s, \hat{\mathbf{y}}) = \sum_i \min(\mathbf{y}_i^s, \hat{\mathbf{y}}_i). \quad (5)$$

In $\mathcal{L}_{\text{KLD}}$, $\mathcal{L}_{\text{CC}}$ and $\mathcal{L}_{\text{SIM}}$, $\mathbf{y}^s$, and $\hat{\mathbf{y}}$ are normalized so that $\sum_i \mathbf{y}_i^s = \sum_i \hat{\mathbf{y}}_i = 1$. According to our empirical studies, the weights λ_1, λ_2, λ_3, and λ_4 of the combined loss function are set to -1, 10, -2, and -1 respectively for balancing the impact of each sub-loss.

3. EXPERIMENTS

3.1. Datasets

- **SALICON** [17] is one of the largest human visual attention datasets and has been widely used to pre-train saliency models, which contains 10,000 and 5,000 natural images and visual attention data for training and validation, respectively. Instead of using eye-trackers, the visual attention is collected by mouse-clicking. In this study, this dataset is used to explore the impact of pre-training models using a large-scale eye-movement dataset of natural images on predicting saliency for diagnostic images.
- **Mammogram eye-tracking dataset** consists of 196 mediolateral oblique (MLO) view mammogram images from 98 anonymous cases and the eye-tracking data of 10 expert radiologists. The details of this dataset can be found in [7]. This dataset is used to investigate the performance of the models in predicting the visual attention of radiologists.

3.2. Evaluation Metrics

Various evaluation metrics have been used to evaluate the agreement of predicted saliency maps with ground truth. Based on their characteristics [18], five widely used evaluation metrics, including CC, SIM, KLD, NSS, and area under ROC curve (AUC), were used to evaluate the models in this study for a comprehensive and fair comparison.

3.3. Experimental Settings

Two training stages were used to train our models, including pre-training and fine-tuning, and the optimal models were determined when the loss values on the validation set in five consecutive epochs were consistently higher than the recorded minimum loss. In the pre-training stage, the deep encoder was initialized by the pre-trained parameters for ImageNet [19], and then the models were trained on the SALICON dataset. In the fine-tuning stage, k-fold Cross-Validation ($k = 7$) was applied to obtain comprehensive results. Specifically, the eye-tracking mammogram dataset was divided into seven non-overlapping subsets, and each subset contained 28 images from 14 cases. To eliminate randomness, each test set corresponded to a fixed validation set and a training set. In each fine-tuning and testing, one subset was kept as a test set, one as a validation set, and the remaining five subsets were used as a training set jointly. The optimal models were obtained using the same method as the pre-training stage and then tested on the corresponding test set. The report results were the average performance on the seven tests. In both training stages, Adam [20] was adopted as the learning strategy. The initial learning rates for pre-training and fine-tuning were set to 2×10^{-5} and 2×10^{-4}, respectively, and they were reduced by multiplying by a factor of 0.1 every 3 epochs. When the model was trained directly on the mammogram eye-tracking dataset without loading the pre-trained parameters of SALICON, the initial learning rate was set to 2×10^{-4} and reduced by multiplying by a factor of 0.1 every 12 epochs to ensure adequate training. Besides, to save computational resources, all input images were resized to the size of 384×288 pixels.

3.4. Results & Discussion

3.4.1. Impact of pre-training with a large-scale eye-movement dataset of natural images

Previous studies have shown that pre-training deep learning models on large-scale natural image datasets is beneficial for deep learning-based algorithms on diagnostic image-related tasks [8]. Accordingly, we use the parameters pre-trained on ImageNet to initialize the deep encoder. Similarly, there are also large-scale natural image saliency datasets, i.e., SALICON, in the eye-movement domain. However, there are significant differences between SALICON and the medical image eye-tracking dataset in terms of targeting tasks (free-viewing vs. medical diagnosis), types of visual selection (bottom-up vs. top-down), and data collection methods (mouse-clicking vs. eye-tracking). Therefore, it is necessary to investigate the effect of pre-training models on SALICON in predicting the visual attention of radiologists. The impact of pre-training is shown in Table 1. It shows pre-training on SALICON can improve the model's performance in multiple metrics on the prediction of mammogram images' saliency. Since diagnostic images with expert annotations are difficult to obtain in large quantities, it would be useful to pre-train saliency models with large-scale eye-movement datasets of natural images. In the subsequent experiments, all models are initialized with the parameters pre-trained on SALICON.

3.4.2. Contributions of high-resolution image representations and shallow encoder

The contributions of deep high-resolution image representations and the shallow encoder can be seen from Table 2. Because the *HRNet* also provides features with the spatial size of $\frac{w}{32} \times \frac{h}{32}$ at the end of its final stage, we use it to simulate the output of an encoder with heavily downsampling, which is denoted as *Deep encoder* $_{\frac{1}{32}}$. In this variant, the spatial size of the outputs from the encoder is $384 \times \frac{w}{32} \times \frac{h}{32}$. Three blocks

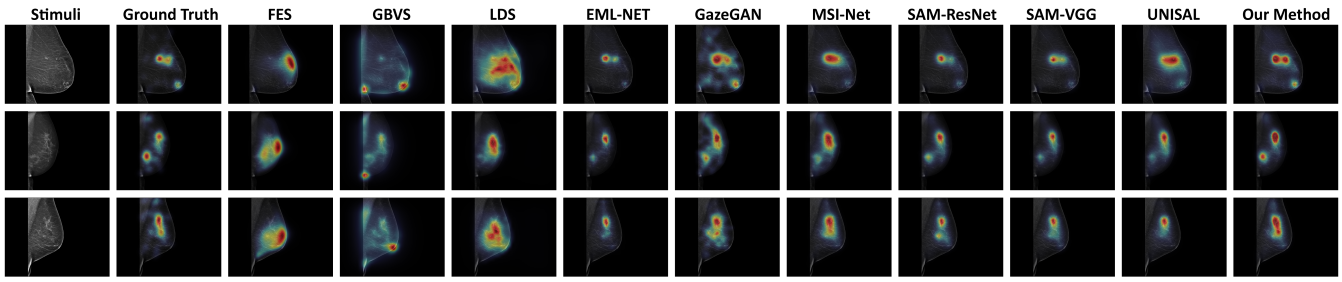

Fig. 2. Illustrations of estimating the visual attention of radiologists performing breast screening using computational models. The first and second columns on the left present the mammograms and the ground truth visual attention of radiologists.

Table 1. The performance on prediction radiologists' visual attention of the proposed model pre-trained on or not on SALICON.

Pre-training conditions	CC ↑	SIM ↑	KLD ↓	NSS ↑	AUC ↑
without SALICON	0.8824	0.7539	0.2760	2.8956	0.9423
with SALICON	0.9015	0.7771	0.2433	2.9912	0.9444

Table 2. The performance of model variants with different encoders in the encoding stage.

Adopted encoder	CC ↑	SIM ↑	KLD ↓	NSS ↑	AUC ↑
Deep encoder$_{\frac{1}{32}}$ (Baseline)	0.8951	0.7722	0.2740	2.9582	0.9437
Deep encoder$_{\frac{1}{4}}$	0.8974	0.7736	0.2457	2.9702	0.9439
Shallow + Deep encoder$_{\frac{1}{4}}$	0.9015	0.7771	0.2433	2.9912	0.9444

Table 3. Comparison results of saliency models. The **Bold** fonts indicate the best score. The models are sorted by their CC scores, which is based on the suggestions of [18].

MODEL	CC ↑	SIM ↑	KLD ↓	NSS ↑	AUC ↑
FES [21]	0.6437	0.5819	1.2958	2.0345	0.9084
GBVS [22]	0.6533	0.5245	0.7188	2.0798	0.9030
LDS [23]	0.7496	0.6270	0.6447	2.3751	0.9157
GazeGAN [13]	0.8543	0.7338	0.5250	2.7781	0.9383
UNISAL [14]	0.8680	0.7476	0.3011	2.8214	0.9399
SAM-ResNet [11]	0.8855	0.7618	0.3052	2.9095	0.9417
MSI-Net [12]	0.8871	0.7636	0.2609	2.8867	0.9418
SAM-VGG [11]	0.8908	0.7687	0.3297	2.9503	0.9426
EML-NET [15]	0.8909	0.7668	0.2680	2.9876	0.9435
Our Method	**0.9015**	**0.7771**	**0.2433**	**2.9912**	**0.9444**

that have a similar structure as decoder block *D_1* and *D_2* (see in Fig. 1) are added at the connection between encoder and decoder to reduce the channels (384→192→96→48) and restore the spatial size of the feature maps ($\frac{1}{32} \rightarrow \frac{1}{16} \rightarrow \frac{1}{8} \rightarrow \frac{1}{4}$). Correspondingly, the model that only adopts a deep encoder but without a shallow encoder is denoted as *Deep encoder*$_{\frac{1}{4}}$. We can see that the Deep encoder$_{\frac{1}{4}}$ outperforms the Deep encoder$_{\frac{1}{32}}$ on all metrics. This illustrates that employing higher spatial size features is beneficial for estimating the saliency of diagnostic images. Furthermore, by adding the shallow encoder to provide higher spatial size features at a shallow depth, the performance of *Shallow + Deep encoder*$_{\frac{1}{4}}$ is further boosted on these five metrics.

3.4.3. Comparison of other saliency models

In order to further validate the proposed method, nine visual saliency prediction models were selected to compare their performance with the proposed model. Among these models, three are traditional saliency models, including FES [21], GBVS [22], and LDS [23], and six are deep learning-based models, including GazeGAN [13], UNISAL [14], SAM-ResNet [11], SAM-VGG [11], MSI-Net [12], and EML-NET [15]. These models represent state-of-the-art traditional or deep learning-based models on the MIT300 benchmark, and their source codes and pre-trained models are publicly available. For a fair comparison, the deep learning-based saliency model was first initialized by parameters pre-trained on the SALICON and fine-tuned appropriately on the mammogram dataset with the same k-fold Cross-Validation ($k = 7$) strategy as the proposed model, and then reported the average results on seven subsets. The comparison results can be seen in Table 3. Our method achieved the best results across these five metrics. In addition, examples of these models' prediction results are shown in Fig. 2. These results imply that our method provides superior visual saliency prediction in clinical practice.

4. CONCLUSION

In this paper, we proposed a method that adopts high-resolution image representations to predict radiologists' visual attention during mammogram interpretation. The proposed method achieved state-of-the-art performance on multiple metrics on a large-scale mammogram eye-movement dataset. We also demonstrated the superiority of pre-training on a benchmark eye-movement dataset of natural images and using higher-resolution representations to estimate the saliency of diagnostic images. Future work includes the application of the model on other radiological examinations such as chest radiographs.

5. REFERENCES

[1] Y. Li, H. Cao, C. M. Allen, X. Wang, S. Erdelez, and C. Shyu, "Computational modeling of human reasoning processes for interpretable visual knowledge: a case study with radiographers," *Sci. Rep.*, vol. 10, no. 1, pp. 21620, 12 2020.

[2] O. B. Ahmed, F. Christine, A. Julian, and M. Paccalin, "Visual saliency for medical imaging and computer-aided diagnosis," in *Neurological Disorders and Imaging Physics*, vol. 3. IOP Publishing, 2019.

[3] S. Banerjee, S. Mitra, and B. U. Shankar, "Automated 3d segmentation of brain tumor using visual saliency," *Inf. Sci.*, vol. 424, no. C, pp. 337–353, Jan. 2018.

[4] H. Fan, F. Xie, Y. Li, Z. Jiang, and J. Liu, "Automatic segmentation of dermoscopy images using saliency combined with otsu threshold," *Comput. Biol. Med.*, vol. 85, pp. 75–85, 2017.

[5] P. K. Sran, S. Gupta, and S. Singh, "Visual saliency models applied to ROI detection for brain MR images: A critical appraisal and future prospects," *SN COMPUT. SCI.*, vol. 2, no. 3, pp. 208, 2021.

[6] U. Engelke, A. Maeder, and H. Zepernick, "Analysing inter-observer saliency variations in task-free viewing of natural images," in *2010 IEEE Int. Conf. Image Process.*, 2010, pp. 1085–1088.

[7] J. Lou, X. Zhao, P. Young, R. White, and H. Liu, "Study of saccadic eye movements in diagnostic imaging," in *2021 IEEE Int. Conf. Image Process.*, 2021, pp. 1474–1478.

[8] A. Esteva, A. Robicquet, B. Ramsundar, V. Kuleshov, M. DePristo, K. Chou, C. Cui, G. Corrado, S. Thrun, and J. Dean, "A guide to deep learning in healthcare," *Nat. Med.*, vol. 25, pp. 24–29, 2019.

[9] P. Mukherjee, M. Zhou, E. Lee, A. Schicht, Y. Balagurunathan, S. Napel, R. Gillies, S. Wong, A. Thieme, A. Leung, and O. Gevaert, "A shallow convolutional neural network predicts prognosis of lung cancer patients in multi-institutional computed tomography image datasets," *Nat Mach Intell*, vol. 2, no. 5, pp. 274–282, 2020.

[10] J. Wang, K. Sun, T. Cheng, B. Jiang, C. Deng, Y. Zhao, D. Liu, Y. Mu, M. Tan, X. Wang, W. Liu, and B. Xiao, "Deep high-resolution representation learning for visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 43, no. 10, pp. 3349–3364, 2021.

[11] M. Cornia, L. Baraldi, G. Serra, and R. Cucchiara, "Predicting human eye fixations via an LSTM-based saliency attentive model," *IEEE Trans. Image Process*, vol. 27, no. 10, pp. 5142–5154, 2018.

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

[13] 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 Trans. Image Process*, vol. 29, pp. 2287–2300, 2020.

[14] R. Droste, J. Jiao, and J. A. Noble, "Unified Image and Video Saliency Modeling," in *Proc. 16th Eur. Conf. Comput. Vis.*, 2020, vol. 12350, pp. 419–435.

[15] S. Jia and N. D.B. Bruce, "EML-NET: An expandable multi-layer network for saliency prediction," *Image Vis. Comput*, vol. 95, pp. 103887, 2020.

[16] Z. Liu, H. Mao, C. Wu, C. Feichtenhofer, T. Darrell, and S. Xie, "A ConvNet for the 2020s," *arXiv preprint arXiv:2201.03545*, 2022.

[17] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *2015 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1072–1080.

[18] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 3, pp. 740–757, 2019.

[19] J. Deng, W. Dong, R. Socher, L. Li, Kai L., and Li F., "ImageNet: A large-scale hierarchical image database," in *2009 IEEE Conf. Comput. Vis. Pattern Recognit.*, 2009, pp. 248–255.

[20] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR 2015*, 2015.

[21] H. R. Tavakoli, E. Rahtu, and J. Heikkilä, "Fast and Efficient Saliency Detection Using Sparse Sampling and Kernel Density Estimation," in *Image Anal.*, 2011, pp. 666–675.

[22] J. Harel, C. Koch, and P. Perona, "Graph-Based Visual Saliency," in *Proc. of the 19th Int. Conf. on Neural Inf. Process. Syst.*, 2006, p. 545–552.

[23] S. Fang, J. Li, Y. Tian, T. Huang, and X. Chen, "Learning Discriminative Subspaces on Random Contrasts for Image Saliency Analysis," *IEEE Trans. Neural Networks Learn. Syst.*, vol. 28, no. 5, pp. 1095–1108, 2017.