A Light-Weight Human Eye Fixation Solution for Smartphone Applications

Sudha Velusamy, Rakesh Radarapu, Anandavardhan Hegde, Narayan Kothari
Samsung R&D Institute, Bangalore, India
{sudha.v, rakesh.r77, hegde.ana, k.narayan}@samsung.com

Abstract

*A wide range of human eye fixation prediction algorithms have been presented in the research with the advent of deep learning. However, to generate better prediction outcomes, these methods are becoming increasingly complicated. In this study, we present a lightweight human eye fixation prediction network that is based on a low-complexity representation learning network and can handle a variety of real-world data. The method includes a simplified multi-level feature extraction network with an emphasis on channel and spatial attention mechanism. We investigate the effectiveness of the present technique in predicting eye fixation maps on a collection of challenging images from the SALICON and MIT*1003 *datasets. A comprehensive qualitative and quantitative evaluation revealed that the network could learn and capture spatial and semantic information in a scene effectively, resulting in a higher hit rate and fewer false positives in comparison with the competing solutions. The approach is implemented on Samsung Galaxy S*23 *with SnapDragon-SM*8550 *mobile platform given its short inference time of* $1.4ms$ *and low model complexity.*

1. Introduction

Visual scene interpretation is a crucial research area that is used in several applications such as video monitoring, robot navigation, computer vision, and so on. Human eye fixation detection, the problem of locating points or image regions that engage human observers' attention upon first sight, is one of the critical subjects under visual scene understanding for its specific use cases in object detection and tracking, image design, image retargeting, and so on. According to eye fixation studies, an interesting 'visual stimulus' in a scene stimulates a section of the human eye retina to process complicated information. When it comes to visual stimuli, all features may be classified as either low-level or high-level. Color contrast, orientation, intensity, positioning, and boundaries are a few instances of low level characteristics. Faces, objects, and text are some examples of high-level features. Fig. 1 presents a few example images

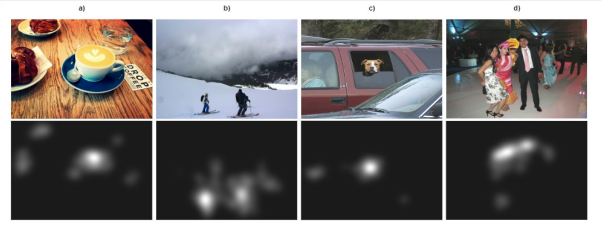

Figure 1. Sample Images and Ground Truth Fixation Maps

from well-known Fixation prediciton datasets with ground truth fixation maps: SALICON [12] and MIT1003 [13].

Early research into eye fixation focused on developing algorithms that employed low-level handcrafted properties. Yet, given the variety of factors that define visual saliency, designing approaches that successfully incorporate all such features individually can be challenging and tedious. In recent years, deep neural networks have proven to be very effective in improving the quality of eye fixation prediction and salient object detection. Several more encoder-decoder-based deep architectures have been proven to increase detection accuracies by emphasising multi-level feature representations and recurrent objectness refinement techniques to integrate both low and high-level characteristics. Existing state-of-the-art models are usually complex and time-consuming to infer. Nevertheless, for real-time use scenarios, increasing parameter overheads and processing complexity have become a bottleneck. Furthermore, for real-time deployments, a balanced network architecture that accommodates wide fluctuations in real-world data, with better detection performance and reduced running costs is required.

EFNet is proposed to handle highly variable real-world data effectively. This network was created primarily with mobile devices in mind, and it delivers impressive accuracy on real-world data samples. A simplified multi-level feature extraction network is a component of this proposed architecture to provide context-rich representations. To handle the diverse data distributions, we employ a channel and spatial attention, with an emphasis on discriminative representations. Using stage-wise fixation prediction on

a challenging set of images from the benchmark datasets, $SALICON$ and $MIT1003$, the proposed technique effectively captures a wide variety of visual context information.The quantitative and qualitative results presented in Section 4 demonstrate the robustness of the proposed EFNet as compared to the state-of-the-art methods. The solution is ideal for low-power devices like mobile phones for its faster inference speed and minimal model complexity

The primary contributions of the proposed solution are summarized as follows:

1. A comprehensive study and comparison of the-state-of-art methods.
2. A scalable, lightweight solution for fixation prediction, that is well-suited for low-power devices.

2. Related Work

Many computational models have already been presented to forecast human eye fixations and saliency maps. There is a strong association between the eye fixation and saliency maps. The former predicts sparse human eye fixation spots in a picture, while the latter uses a two-dimensional topographically organised map to precisely detect the whole attentive object regions. Most early algorithms for fixation prediction were based on traditional computer vision techniques that produced pixel-level properties, such spectral residue, global context data, etc. Moreover, it has been shown that the performance of detection may be enhanced by the use of hand-crafted picture priors [2, 3, 3, 10, 10, 19, 19, 22].

Deep models are now being employed in this field of study since earlier approaches that were more concerned with pixel-level visual features were unable to adequately collect the semantic data required for such complex tasks. The state-of-the-art in fixation prediction has greatly advanced because to the establishment of deep neural networks (DNNs) and the availability of large-scale saliency data sets. Using a 3 layer network and DNN, a preliminary attempt to model saliency is shown in [20]. Following that, Kuemmerer et al [15] proposed a transfer learning approach that creates saliency maps using pre-existing networks trained for object recognition tasks. Subsequently, it became clear that models built using Fully Convolutional Networks (FCN) were more effective and successful in predicting saliency. A novel saliency prediction model was put out by Dodge et al [7] that takes into account both local information produced by a DNN and scene-wide semantic data.

Salient and non-salient areas at various sizes were employed in network construction for eye fixation prediction by Liu et al [17]. Wang et al [21] developed multi-level supervision in the convolutional layers with different receptive fields and a skip-layer network topology to predict human eye fixation. By using location-based convolution filters, the approach described by DeepFix [14] enables the network to take advantage of location-dependent patterns. In a different research, SALICON [12], saliency is predicted using a multi-stream approach and a network objective function that is customized for saliency.

In conclusion, the majority of the approaches described above focus on deep network variations to capture the representation of several layers of features, producing heavier models that are still unable to handle the majority of data variations in real-world samples.

3. Proposed Method

The architecture of the proposed fixation prediction method, EfficientFixationNetwork, referred to as EFNet, is shown in Figure 2. EFNet is composed of two main parts: i) Deep Feature Extraction Module, to extract deep feature representations of the image; ii) Feature Attention Module, to emphasize significant and context-rich representations. In this section, each of these stages is explained in detail.

3.1. Deep Feature Extraction

With the goal to develop an effective and simple model for fixation prediction, we employ EfficientNet [18] as a backbone network to acquire deep feature representations, by drawing inspiration from TRACER [16]. EfficientNet offers greater learning capabilities and is compact as compared to other models like ResNet and VGG. We experimentally choose feature maps from three different stages of the CNN: 3, 5, and 7 which contain 40, 112, and 320 channels, respectively. These are represented in the model architecture as F_3, F_5, F_7. The feature maps are reduced to $\hat{F}_3$, $\hat{F}_5$ and $\hat{F}_7$ of sizes 32, 64 and 128, by processing through multi-kernel based receptive field blocks, which have a set of $k \times 1$ and $1 \times k$ convolutions. $\hat{F}_5$, $\hat{F}_7$ are upsampled by scale factors 2, 4, respectively and concatenated to $\hat{F}_3$, along the channel axis, giving multi-level feature maps.

For mobile applications, we employ EfficientNet-lite1 model for deep feature extraction. This lite version is tailored from the original EfficientNet and is well supported by mobile accelerators. Some modifications include, removal of squeeze-and-excitation networks, replacement of all swish activations with RELU6 to improve the quality of post-training quantization.

3.2. Feature Attention Module:

The Feature Attention Module receives the multi-level feature maps from the Deep Feature Extraction Module as input. As these feature maps capture information at different levels, we employ channel and spatial attention blocks. Channel attention is used to emphasize the significant channels from the input feature representations. The spatial in-

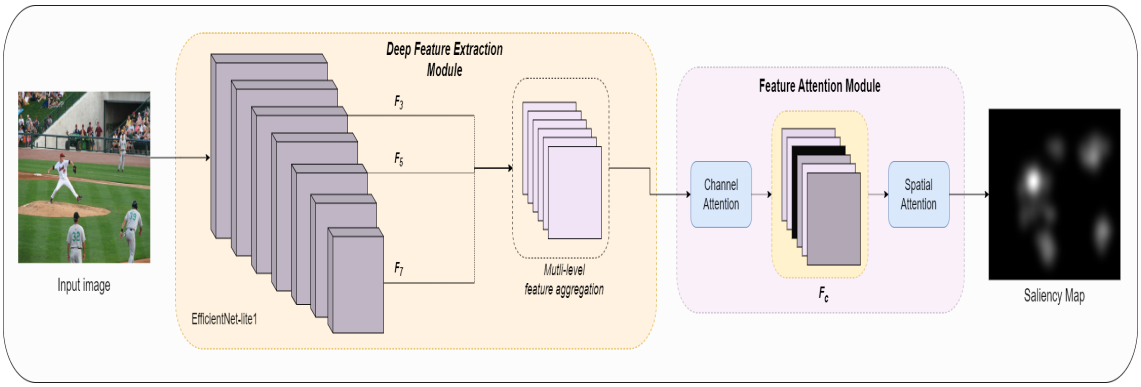

Figure 2. The Proposed EFNet Architecture for Eye Fixation Map Prediction

formation is mean-pooled globally to obtain a representative value $\tilde{X}$, for each feature map. Self-attention is applied to generate the attention map α_c using the descriptor $\tilde{X}$ as shown in Eq. 1.

$$\alpha_c = \sigma\left(\frac{\exp(\mathcal{F}_q(\tilde{X})\mathcal{F}_k(\tilde{X})^T)}{\sum exp(\mathcal{F}_q(\tilde{X})\mathcal{F}_k(\tilde{X})^T)}\mathcal{F}_v(\tilde{X})\right) \tag{1}$$

where $\mathcal{F}(\cdot)$ is a convolution operation using a 1×1 kernel. The final representation of channel attention is given by

$$X_c = X * \alpha_c + X \tag{2}$$

Supporting channel attention, spatial attention is employed to focus on the feature maps' informative regions. The inter-spatial relationship of features is captured using self-attention, and the input data is reduced to a single output feature X_s.

$$X_s = \left(\frac{\exp(\mathcal{F}_q(X_c)\mathcal{F}_k(X_c)^T)}{\sum exp(\mathcal{F}_q(X_c)\mathcal{F}_k(X_c)^T)}\mathcal{F}_v(X_c)\right) + \mathcal{F}_v(X_c) \tag{3}$$

The final saliency map, denoted as S_m, is generated by passing X_s from the preceding stage through a sigmoid layer.

3.3. Adaptive Pixel Intensity Loss

To create the loss function, similar to TRACER [16], we combine the binary cross entropy (BCE), intersection over union (IoU), and L1 loss functions. We observed that, even though Salient Object Detection was its original use, Fixation Prediction can benefit from it as well. It effectively highlights the most salient region in comparison to the surrounding area by taking into consideration the pixel intensity w.

$$w_{ij} = (1-\lambda)\sum_{k \in K}\left|\frac{\sum_{h,w \in A_{ij}} y^k_{hw}}{\sum_{h,w \in A_{ij}} 1} - y_{ij}\right| y_{ij} \tag{4}$$

In Eq. 4, K denotes the kernel size and (h, w) represents the pixels around the target pixel A_{ij} within the kernel. λ is a penalty term set to 0.5 and kernel size $K \in \{3, 15, 31\}$.

In BCE loss, the pixel intensity w is used to help the network zero in on the size of the salient areas. The adaptive BCE loss is shown in Eq. 5, where y and $\hat{y}$ indicate the label and predicted probability of the binary class c.

$$\mathcal{L}^a_{BCE} = -\frac{\sum_i^H\sum_j^W(1+w_{ij})\sum_{c=0}^1(y_c\log(\hat{y_c}) + (1-y_c)\log(1-\hat{y_c}))}{\sum_i^H\sum_j^W(1.5+w_{ij})} \tag{5}$$

Equation 6 shows a modified IoU loss that places more emphasis on the bright pixels than the other pixels.

$$\mathcal{L}^a_{IoU} = 1 - \left(\frac{\sum_i^H\sum_j^W(y_{ij}\hat{y}_{ij})(1+w_{ij})}{\sum_i^H\sum_j^W(y_{ij}+\hat{y}_{ij}-y_{ij}\hat{y}_{ij})(1+w_{ij})}\right) \tag{6}$$

We apply the pixel intensity w to L1 loss, as shown in Eq. 7. When calculating the deviation from the ground truth, this aids in the differentiation of significant pixels.

$$\mathcal{L}^a_{L1} = \frac{\sum_i^H\sum_j^W|y_{ij}-\hat{y}_{ij}|(1+w_{ij})}{H*W\sum_i^H\sum_j^W w_{ij}} \tag{7}$$

The final loss function, referred to as Adaptive Pixel Intensity loss, is created by combining the above 3 loss functions as shown below,

$$\mathcal{L}_{API}(y,\hat{y}) = \mathcal{L}^a_{BCE}(y,\hat{y}) + \mathcal{L}^a_{IoU}(y,\hat{y}) + \mathcal{L}^a_{L1}(y,\hat{y}) \tag{8}$$

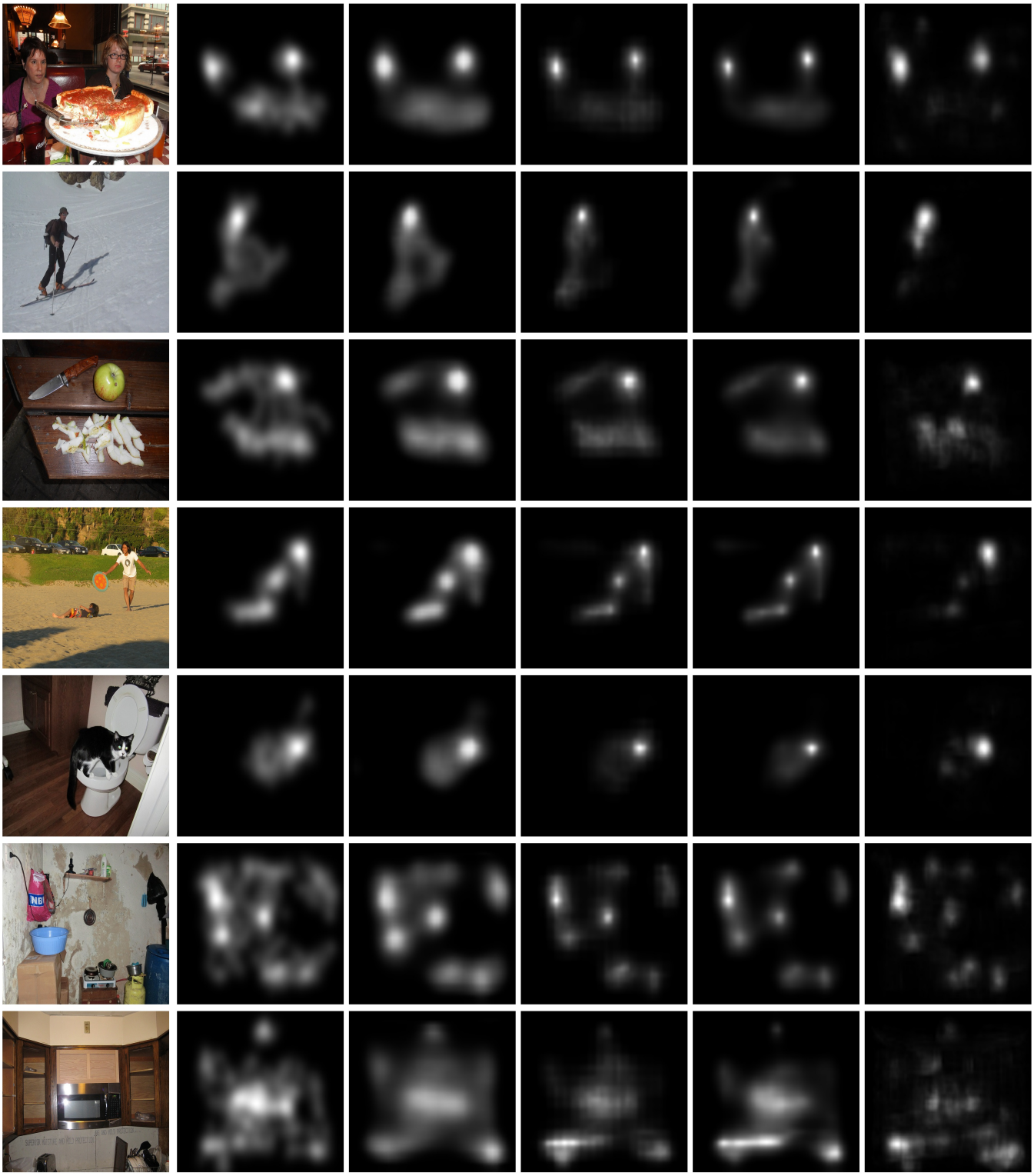

Figure 3. Qualitative Comparison of Fixation Maps: EFNet Vs Competitor Methods

Table 1. Quantitative Comparison of Detection Performance: EPNet Vs Competitor Methods

Method	AUC ↑	sAUC ↑	NSS ↑	CC ↑	Sim ↑	KLD ↓	MB ↓
SalFBNet [6]	**0.868**	0.740	1.952	0.892	0.772	**0.236**	23.4
FB Net [5]	0.843	0.706	1.687	0.785	0.694	0.708	**4.7**
MD-SEM [9]	0.864	0.746	**2.058**	0.868	0.774	0.568	-
EML Net [11]	0.866	0.746	2.050	0.886	0.780	0.520	180.2
UNISAL [8]	0.864	0.739	1.952	0.879	0.775	-	14.7
GazeGAN [1]	0.864	0.736	1.899	0.879	0.773	0.376	879.2
ML Net [4]	0.866	**0.768**	-	0.743	-	-	58.9
EFNet-lite	0.860	0.740	1.942	**0.892**	**0.782**	0.870	**15.0**
EFPNet-E1	0.861	0.742	1.973	0.900	0.790	0.907	21.6

4. Experimentations and Results

4.1. Datasets and Metrics

For our training and assessment purposes, we employ the datasets $SALICON$ and $MIT1003$, respectively. SALICON (SALIency in CONtext) is a large dataset for selective attention that includes $20K$ mouse-tracking annotated pictures. The dataset was constructed using samples from the well-known MS COCO dataset. $10K$, $5K$, and $5K$ pictures from the $20K$ photos are used as training, testing, and validation sets, respectively. Images from the Flicker and LabelMe datasets totaling 1003 are included in the $MIT1003$ dataset. It is based on eye-tracking data from fifteen people who watched the photos at their own discretion.

We employ Similarity (SIM), Kullback-Leibler divergence (KLD), Area under ROC Curve (AUC), Shuffled AUC (sAUC), Normalized Scanpath Saliency (NSS), and Pearson's Correlation Coefficient (CC) as metrics for our validation and comparisons with competing algorithms based on the extensive study reference metrics.

4.2. Implementation Details

MIT1003 and SALICON datasets are used to train and assess the presented approach. We follow the SALICON recommended partitioning for validation and testing. With a maximum of 50 epochs, the training batch size is set at 32. For each epoch, the Adam optimizer was employed with a learning rate of 5×10^{-5} and a weight decay of 10^{-4}. The proposed model is benchmarked against various competitive approaches using a TITAN-X GPU.

EFNet-E1, based on EfficientNet-B1 was developed as the primary version for the task. Then a mobile compatible variant called EFNet-lite was developed based on EfficientNet-lite1. EFNet-lite is only $9MB$ and operates at $90fps$. Due to the restricted support for floating-point operations on mobile devices, the lite version of the model is then quantized to 8 bits. On the accelerated mobile platform, this quantized variant runs at 715 fps.

4.3. Results

4.3.1 Comparative Results

To demonstrate the efficacy of our current method, we conducted quantitative and qualitative analyses on the $SALICON$ test set of $5,000$ samples. The quantitative findings from submitting the predicted fixation maps to the challenge system [1] are shown in Table 1. The suggested method's improved performance, particularly in terms of CC and Similarity measures, is demonstrated through the performance comparisons with recent popular state-of-the-art approaches including EML-Res, EML-Nas [11], and MLNet [4].

In order to show how the recommended technique can handle challenging real-world samples, we present the comparative fixation maps in Fig. 3. A collection of unseen images from the SALICON dataset are selected for comparison in order to account for subjective observation. The high-quality fixation maps predicted by the present method are found to be comparatively consistent and smooth when compared to the ground truth maps. The samples in the figure clearly show the efficacy of the proposed approach owing to the improved hit rate and lowered false positives.

5. Conclusions

In this paper, we presented a lightweight human eye fixation prediction network that is robust to real-world data variations with comparable better accuracy with the latest state-of-the-art techniques. The improved inference time and low model complexity of the proposed method is highly suitable for solution deployment in low-power devices like smart phones.

References

[1] Zhaohui Che, Ali Borji, Guangtao Zhai, Xiongkuo Min, Guodong Guo, and Patrick Le Callet. How is gaze influ-

[1] http://salicon.net/challenge-2017/

enced by image transformations? dataset and model. *IEEE Transactions on Image Processing*, 2019. 5

[2] Ming-Ming Cheng, Niloy Jyoti Mitra, Xiaolei Huang, Philip H. S. Torr, and Shimin Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2015. 2

[3] Ming-Ming Cheng, Jonathan Warrell, Wen-Yan Lin, Shuai Zheng, Vibhav Vineet, and Nigel Crook. Efficient salient region detection with soft image abstraction. pages 1529–1536, 2013. 2

[4] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A deep multi-level network for saliency prediction. 2016. 5

[5] Guanqun Ding, Nevrez Imamoglu, Ali Caglayan, Masahiro Murakawa, and Ryosuke Nakamura. Fbnet: Feedback-recursive cnn for saliency detection. *2021 17th Intl. Conf. on Machine Vision and Applications (MVA)*, 2021. 5

[6] Guanqun Ding, Nevrez İmamoğlu, Ali Caglayan, Masahiro Murakawa, and Ryosuke Nakamura. Salfbnet: Learning pseudo-saliency distribution via feedback convolutional networks. *Image and Vision Computing*, 120:104395, 2022. 5

[7] Samuel F. Dodge and Lina Karam. Visual saliency prediction using a mixture of deep neural networks. *IEEE Transactions on Image Processing*, 27, 2017. 2

[8] Richard Droste, Jianbo Jiao, and Julia Alison Noble. Unified image and video saliency modeling. *ArXiv*, abs/2003.05477, 2020. 5

[9] Camilo Luciano Fosco, Anelise Newman, Patr Sukhum, Yun Bin Zhang, Nanxuan Zhao, Aude Oliva, Zoya Bylinskii, and Hong Kong. How much time do you have? modeling multi-duration saliency. *2020 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[10] Bo Fu, Yong-Gang Jin, Fan Wang, and Xiao-Peng Hu. Prior fusion based salient object detection. 2014. 2

[11] Sen Jia. Eml-net: An expandable multi-layer network for saliency prediction. *ArXiv*, 2018. 5

[12] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. Salicon: Saliency in context. In *2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015. 1, 2

[13] Tilke Judd, Krista Ehinger, Frédo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th Intl. Conf. on Computer Vision*, 2009. 1

[14] Srinivas S. S. Kruthiventi, Kumar Ayush, and R. Venkatesh Babu. Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 2015. 2

[15] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition. *ArXiv*, 2016. 2

[16] Min Seok Lee, WooSeok Shin, and Sung Won Han. TRACER: extreme attention guided salient object tracing network. *CoRR*, abs/2112.07380, 2021. 2, 3

[17] Nian Liu, Junwei Han, Dingwen Zhang, Shifeng Wen, and Tianming Liu. Predicting eye fixations using convolutional neural networks. *2015 IEEE Conf. on Computer Vision and Pattern Recog. (CVPR)*, 2015. 2

[18] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019. 2

[19] Na Tong, Huchuan Lu, Xiang Ruan, and Ming-Hsuan Yang. Salient object detection via bootstrap learning. pages 1884–1892, 2015. 2

[20] Eleonora Vig, Michael Dorr, and David D. Cox. Large-scale optimization of hierarchical features for saliency prediction in natural images. *2014 IEEE Conf. on Computer Vision and Pattern Recognition*, 2014. 2

[21] Wenguan Wang and Jianbing Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27, 2017. 2

[22] Wangjiang Zhu, Shuang Liang, Yichen Wei, and Jian Sun. Saliency optimization from robust background detection. page 2814–2821, 2014. 2