FixPNet: Attention Guided Fixation Map Prediction With Explicit Image Priors

Rakesh Radarapu Sudha Velusamy Anandavardhana Hegde Narayan Kothari

Samsung R&D Institute, Bangalore, India
{rakesh.r77, sudha.v, hegde.ana, k.narayan}@samsung.com

Abstract

Fixation prediction has been an active area of research in the field of visual scene understanding. To achieve improved prediction results, emerging deep learning solutions are becoming more complex. In this work, we present FixPNet, a novel, and lightweight two-stream fixation prediction network. The proposed architecture is built on attention-driven image priors and a low-complexity representation learning network that can handle a wide variety of real-world data. FixPNet incorporates a simplified multi-level feature extraction network and a parallel stream to derive coarse-level image priors. We examine the significance of image priors by validating on a set of challenging images from the SALICON and $MIT1003$ datasets. Comprehensive qualitative and quantitative evaluation show that the proposed network could learn and capture spatial and semantic information in a scene quite effectively, with a higher hit rate and fewer false positives. The proposed methodology achieves state-of-the-art performance on the SALICON dataset. Given its low inference time and model complexity, FixPNet is ideal for deployment on low-power devices such as mobile phones.

1. Introduction

Visual scene understanding is an important research area that plays a central role in many applications such as video surveillance, robot navigation, visual search, and so on. Human eye fixation detection, the challenge of locating points or image regions that attract human observers' attention, at first sight, is one of the crucial topics under visual scene understanding for its specific use cases in object tracking, image composition, image retargeting, etc. According to research on eye fixation, an interesting region of 'visual stimulus' in a scene triggers a portion of the human eye retina to process complex information. The spectrum of visual stimuli includes low-level features like color contrast, intensity, orientation, position, boundaries, motion, etc., and high-level features like faces, objects, and text. Fig. 1 presents a few example images with ground truth fixation maps from the well-known visual saliency datasets: SALICON [14] and MIT1003 [15].

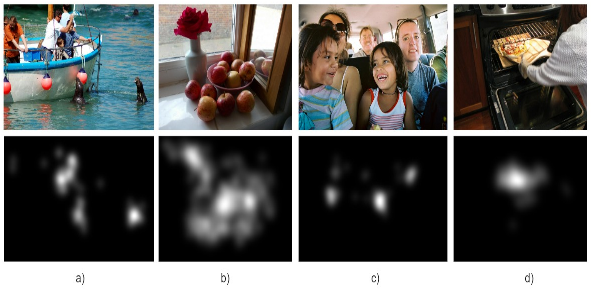

Figure 1. Sample Images and Ground Truth Fixation Maps

Early studies on the subject of eye fixation focused on designing methods using low-level handcrafted features. However, given the range of aspects that define visual saliency, it can be challenging to design methods that effectively combine all such features individually. Deep neural networks have been incredibly successful in improving eye fixation prediction and Salient Object Detection [19] in recent years. Numerous encoder-decoder-based deep architectures that emphasize multi-level feature representations and repeated objectness refinement strategies to incorporate both low and high-level features have been demonstrated to improve detection accuracies. But for real-time use cases, the increased parameter overheads and processing complexity with most methods have turned into a bottleneck. In conclusion, a balanced network design that handles broad variations in real-world data, with improved detection performance and low operating cost is required for real-time deployments.

In order to robustly handle the highly variable real-world data, we present FixPNet, a lightweight fixation prediction network based on attention-guided image priors. The proposed two-stream architecture includes a simplified multi-level feature extraction network to derive context-rich representations and a parallel stream to derive coarse-level image priors that explicitly cover the spectrum of fixation stimuli. We use channel and spatial attention to handle the different data distributions and place emphasis on discriminative representations. Through this two-stream design, the proposed technique effectively captures widely diverse visual context through stage-wise fixation prediction on a challenging set of images from the benchmark datasets, $SALICON$ and $MIT1003$. The quantitative and qualitative results presented in Section 4 demonstrate the robustness of the proposed FixPNet 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. To our knowledge, this is the first work for fixation prediction that studies the significance of adding different priors guided by attention mechanisms to deep feature networks.

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

1. A scalable, lightweight solution for fixation prediction, that is well-suited for low-power devices.
2. A novel two-stream fixation prediction network that exploits both deep learning and traditional visual features.
3. The usage of explicit image priors, which improves the hit rate and reduces the false positives in the predictions.

2. Related Work

To predict human eye fixations and salient object maps, various computational models have been proposed in the past. The eye fixation and salient object maps have a significant correlation; the former predicts sparse human eye fixation locations in an image, while the latter aims to precisely detect the whole attentive object areas using a two-dimensional topographically arranged map.

One of the earliest models for determining center-surround saliency was put forth by Koch et al. [16] and later implemented by Itti et al. [12] with a reasonable amount of success. The majority of the early approaches for fixation prediction were based on conventional computer vision methods that created pixel-level attributes, such as spectral residue, global context information, etc. Additionally, it has been demonstrated that incorporating hand-crafted image priors can improve detection performance [3,4,11,22]. Tong et al [22], Cheng et al [4] quantitatively show the importance of the center prior to modern detection methods. Similarly, the use of boundary, background, contrast, color, and compactness priors are also investigated for their impacts on fixation prediction. For example, the spatial layouts of the objects seen in the images are captured by the background priors as presented by Zhu et al [25]. Cheng et al [3] applied boundary connectivity prior to handling the challenges with background priors. To aid in capturing local and global features, BoFu et al [11] put forth a contrast prior.

Recently, the research in this area has moved to deep models as methods focused on pixel-level visual attributes failed to capture sufficient semantic information, which is crucial for such advanced tasks. With the development of deep neural networks (DNNs) and the availability of large-scale saliency data sets, state-of-the-art in fixation prediction has improved significantly. A first attempt to model saliency using DNN with a 3 layer network is presented by Vig et al [23]. Then Kuemmerer et al [18] suggested a transfer learning method that makes use of already-existing networks trained for object identification tasks to generate saliency maps. Later, it was established that models based on Fully Convolutional Networks (FCN) were more successful and efficient at predicting saliency. Dodge et al [8] proposed a novel saliency prediction model that incorporates global scene semantic information and local information generated by a DNN.

Liu et al [20] used salient and non-salient regions at many scales in network design for eye fixation prediction. In order to predict human eye fixation, Wang et al [24] built multi-level supervision in the convolutional layers with various receptive fields and a skip-layer network structure. The method presented by DeepFix [17] incorporates location-based convolution filters, enabling the network to exploit location-dependent patterns. Another study titled SALICON [14] uses a multi-stream technique to predict saliency with a network objective function that is designed specifically for saliency.

In summary, the majority of the methods outlined above concentrate on deep network variations to capture the representation of several levels of features, resulting in heavier models that are still incapable of handling most data variations in real-world samples.

3. Proposed Method

Figure 2 illustrates the architecture of the proposed fixation prediction method, termed FixPNet. There are three primary components in FixPNet: $i)$ Deep Feature Extraction Module; $ii)$ Prior Generation Module to explicitly cover the spectrum of fixation stimuli; $iii)$ Union Attention Module to emphasize significant and context-rich representations. In this section, each of these stages is explained in detail.

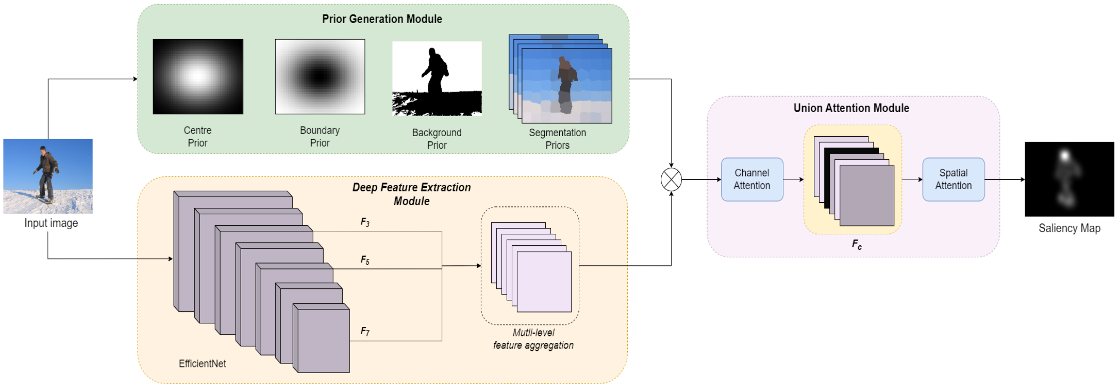

Figure 2. The Proposed FixPNet Architecture for Eye Fixation Map Prediction

3.1. Deep Feature Extraction

We aim to develop an effective and simple model for fixation prediction by drawing inspiration from TRACER [19]. We employ EfficientNet [21] as a backbone network to acquire deep feature representation given its greater learning capabilities and compactness 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.

3.2. Image Priors Generation:

In order to cover the broad spectrum of image stimuli, we choose to generate 4 types of fundamental and complimentary priors: Center, Boundary, Background, and Segmentation. The generation procedure for each of the priors is explained in detail below.

3.2.1 Centre and Boundary Prior:

In order to capture the center bias in the input data, a $2d$-Gaussian map is taken as a Centre Prior. The Gaussian map is generated using Eq. 1 with experimentally selected parameters.

$$(1/2\pi\sigma_x\sigma_y) * \exp^{-[(x-\mu_x^2)/2\sigma_x^2+(y-\mu_y^2)/2\sigma_y^2]} \qquad (1)$$

where σ_x, σ_y denote standard deviations along x, y axes, respectively. μ_x and μ_y denote mean values along x and y axes respectively. The inverse of the center prior calculated above is utilized as the boundary prior.

3.2.2 Background Prior:

Background prior is used to suppress the output saliency map's false detection of salient regions. The steps below are used to generate the background prior.

1. Run edge detection using Sobel Operator.
2. To reduce noise, only retain edges that are above a set threshold.
3. Run contour detection over the edges.
4. Select only significant contours that constitute at least 5% of the total area of the input image.
5. On the resulting image, use the logical NOT operator to produce the final background map.

3.2.3 Segmentation Prior:

One can choose to use any of the most recent, affordable segmentation algorithms to produce coarse-level segmentation maps, as shown in the Fig 2. We decide to create a segmentation map using the well-known clustering-based segmentation algorithm $SLIC$ that is efficient in learning the continuity between salient regions and their extent. By employing multi-level segment maps, wide variations in the input data can be captured. We consider 4 levels of segmentation with the cluster sizes of $5, 25, 45, 65$ to obtain 4 different segment maps.

3.3. Union Attention Module:

The Union Attention Module receives a concatenated input of the feature maps from the Deep Feature Extraction Module and the set of image priors from the Image Prior

Generation Module. As these concatenated feature maps come from different data distributions, we employ channel and spatial attention. To emphasize the significant channels from the input feature representations, channel attention is used. The spatial information is mean-pooled globally to obtain a representative value $\tilde{X}$, for each feature map. We use self-attention and softmax function, with sigmoid function to generate the attention map α_c using the descriptor $\tilde{X}$ as shown in Eq. 2.

$$\alpha_c = \sigma\left(\frac{\exp(\mathcal{F}_q(\tilde{X})\mathcal{F}_k(\tilde{X})^T)}{\sum exp(\mathcal{F}_q(\tilde{X})\mathcal{F}_k(\tilde{X})^T)}\mathcal{F}_v(\tilde{X})\right) \tag{2}$$

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

$$X_c = X * \alpha_c + X \tag{3}$$

In addition to channel attention, spatial attention is used 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.

$$X_s = \left(\frac{\exp(\mathcal{F}_q(X_c)\mathcal{F}_k(X_c)^T)}{\sum exp(\mathcal{F}_q(X_c)\mathcal{F}_k(X_c)^T)}\mathcal{F}_v(X_c)\right) + \mathcal{F}_v(X_c) \tag{4}$$

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

3.4. Adaptive Pixel Intensity Loss

We integrate the binary cross entropy (BCE), Intersection over Union (IoU), and L1 loss functions, much like TRACER [19], to form the loss function. Although its intended purpose was for Salient Object Detection, we learned that it is also useful for Fixation Prediction. By taking into account the pixel intensity w, it helps in effectively highlighting the most salient region in relation to the surrounding area.

$$w_{ij} = (1-\lambda)\sum_{k\,\in\,K}\left|\frac{\sum\limits_{h,w\in A_{ij}} y^k_{hw}}{\sum\limits_{h,w\in A_{ij}} 1} - y_{ij}\right| y_{ij} \tag{5}$$

In Eq. 5, K denotes the kernel size, (h, w) represents the pixels around the target pixel A_{ij} within the kernel and y represents the ground truth label. λ is a penalty term set to 0.5 and kernel size $K \in \{3, 15, 31\}$.

The pixel intensity w is used in BCE loss to enable the network to focus on the extent of the salient regions. Eq. 6 represents the adaptive BCE loss, where y and $\hat{y}$ denote the label and predicted probability of binary class c.

$$\mathcal{L}^a_{BCE} = -\frac{\sum\limits_i^H\sum\limits_j^W(1+w_{ij})\sum\limits_{c=0}^1(y_c\log(\hat{y}_c)+(1-y_c)\log(1-\hat{y}_c))}{\sum\limits_i^H\sum\limits_j^W(1.5+w_{ij})} \tag{6}$$

Equation 7 illustrates adaptive IoU loss, which emphasizes more on the bright pixels more than the other pixels.

$$\mathcal{L}^a_{IoU} = 1 - \left(\frac{\sum\limits_i^H\sum\limits_j^W(y_{ij}\hat{y}_{ij})(1+w_{ij})}{\sum\limits_i^H\sum\limits_j^W(y_{ij}+\hat{y}_{ij}-y_{ij}\hat{y}_{ij})(1+w_{ij})}\right) \tag{7}$$

We apply the pixel intensity w to L1 loss, as shown in Eq. 8. This helps in distinguishing between important pixels when computing the divergence from the ground truth.

$$\mathcal{L}^a_{L1} = \frac{\sum\limits_i^H\sum\limits_j^W|y_{ij}-\hat{y}_{ij}|(1+w_{ij})}{H*W\sum\limits_i^H\sum\limits_j^W w_{ij}} \tag{8}$$

The final loss function referred to as Adaptive Pixel Intensity loss, is taken as a combination of the above 3 loss functions as shown below,

$$\mathcal{L}_{API}(y,\hat{y}) = \mathcal{L}^a_{BCE}(y,\hat{y}) + \mathcal{L}^a_{IoU}(y,\hat{y}) + \mathcal{L}^a_{L1}(y,\hat{y}) \tag{9}$$

4. Experimentations and Results

4.1. Datasets

We use $SALICON$ [14] and $MIT1003$ [15] datasets for our training and evaluation purposes. SALICON (SALIency in CONtext) is a large selective attention dataset that contains $20K$ images with mouse-tracking annotations. The well-known MS COCO dataset's samples were used to build the dataset. From the $20K$ images, $10K$, $5K$, and $5K$ images are used as training, testing, and validation sets, respectively. $MIT1003$ contains 1003 images collected from the Flicker and LabelMe dataset. It is based on eye-tracking data from fifteen subjects who freely viewed the images.

4.2. Evaluation Metrics

Along with the growth of visual saliency literature and its associated datasets, a set of 8 distinct metrics are found to be widely employed. Namely, 1) Area under ROC Curve (AUC); 2) Shuffled AUC (sAUC); 3) Normalized Scanpath Saliency (NSS); 4) Pearson's Correlation Coefficient (CC); 5) Earth Mover's Distance (EMD); 6) Similarity (SIM); 7)

Kullback-Leibler divergence (KLD); 8) Information Gain (IG). Based on the comprehensive study [1], we choose to use AUC, sAUC NSS, CC, Similarity, and KLD as metrics for our validation and comparisons with competing approaches.

4.3. Implementation Details

The proposed model is trained and evaluated using the $SALICON$ and $MIT1003$ datasets. We adhere to the same validation and testing partitions given in SALICON. Employing EfficientNet [21] enables us to develop scaled versions, resulting in gradual improvement in performance at the expense of additional memory and computations. The training batch size is set to 32 with a maximum of 50 epochs. Adam optimizer was used with a learning rate of 5×10^{-5} and a weight decay of 10^{-4} for every epoch. The proposed model was implemented in the Pytorch framework on a TITAN-X GPU and is benchmarked against all other competitive methods.

FixPNet-E2, based on EfficientNet-B2, and FixPNet-E5, based on EfficientNet-B5, along with a lightweight version, named FixPNet-lite based on EfficientNet-lite1 were developed. FixPNet-lite is only 15 MB and operates at $90fps$. It is to be noted that in FixPNet-lite, the feature extraction module has been simplified, which has significantly reduced the size of the model, making it suitable for deployment on low-power devices like mobile phones.

4.4. Results

4.4.1 Comparative Results

We performed both quantitative and qualitative analysis on the $SALICON$ test set of 5000 samples to show the effectiveness of our proposed FixPNet. Table 1 presents the quantitative results obtained by submitting the predicted fixation maps to the challenge system [1]. Performance comparison of the proposed method with the recent and popular state-of-the-art approaches, including EML-Res, EML-Nas [13], and MLNet [5], demonstrates its increased performance, notably in terms of CC and Similarity measures. The choice of the backbone network (EfficientNet variants) results in a trade-off between prediction accuracy and performance cost as shown in Table 1.

We exhibit the comparison fixation maps in Fig. 3 to illustrate how the suggested approach can handle difficult real-world samples. For subjective observation, a set of unseen images from the SALICON dataset are chosen for comparison. When compared to the ground truth maps, the high-quality fixation maps predicted by FixPNet are observed to be relatively smooth and consistent. With the increased hit rate and reduced false positives, the samples in the figure clearly demonstrate the proposed method's effectiveness.

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

Method	**AUC ↑**	**sAUC ↑**	**NSS ↑**	**CC ↑**	**Sim ↑**	**KLD ↓**	**MB ↓**
SalFBNet [7]	**0.868**	0.740	1.952	0.892	0.772	**0.236**	23.4
FB Net [6]	0.843	0.706	1.687	0.785	0.694	0.708	**4.7**
MD-SEM [10]	0.864	0.746	**2.058**	0.868	0.774	0.568	-
EML Net [13]	0.866	0.746	2.050	0.886	0.780	0.520	180.2
UNISAL [9]	0.864	0.739	1.952	0.879	0.775	-	14.7
GazeGAN [2]	0.864	0.736	1.899	0.879	0.773	0.376	879.2
ML Net [5]	0.866	**0.768**	-	0.743	-	-	58.9
FixPNet-lite	0.860	0.740	1.942	**0.892**	**0.782**	0.870	**15.0**
FixPNet-E2	0.861	0.742	1.973	0.900	0.790	0.907	42.9
FixPNet-E5	0.862	0.744	1.991	**0.904**	**0.793**	0.903	120.5

Table 2. Performance of FixPNet with Different Image Priors

FixPNet Configurations	**Sim ↑**	**CC ↑**	**KLD ↓**
Without Priors	0.737	0.871	1.232
Without Segmentation,Background Priors	0.758	0.884	1.012
Without Background Prior	0.756	0.884	1.085
Without Segmentation Prior	0.761	0.888	1.022
With All Priors	**0.781**	**0.901**	**0.893**

4.4.2 Ablation study

We begin with a FixPNet model that comprises all of the prior information and gradually remove each of the four image priors (Center prior, Boundary prior, Segmentation prior, Background prior) to understand the influence of each individual image prior in the fixation prediction task. Table 2 contains the resulting 5 configurations and their stage-by-stage quantitative results. We selected a set of 500 challenging images from the SALICON dataset and compared the performances. The results show improvement by 5.97% in Similarity, 3.4% in CC, and 27.5% in KL Divergence metrics when all 4 priors were taken into account.

We compare fixation maps with and without prior streams in Fig. 4. The fixation maps for the sample set of images highlight the significance of each prior in covering the spectrum of image stimuli through improved hit rates and reduced false positives. We can also note that the FixPNet's predictions with priors are considerably more consistent with the ground truth than the FixPNet's predictions without priors.

5. Conclusions

In this paper, we presented FixPNet, a novel and lightweight fixation prediction network that is robust to real-

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

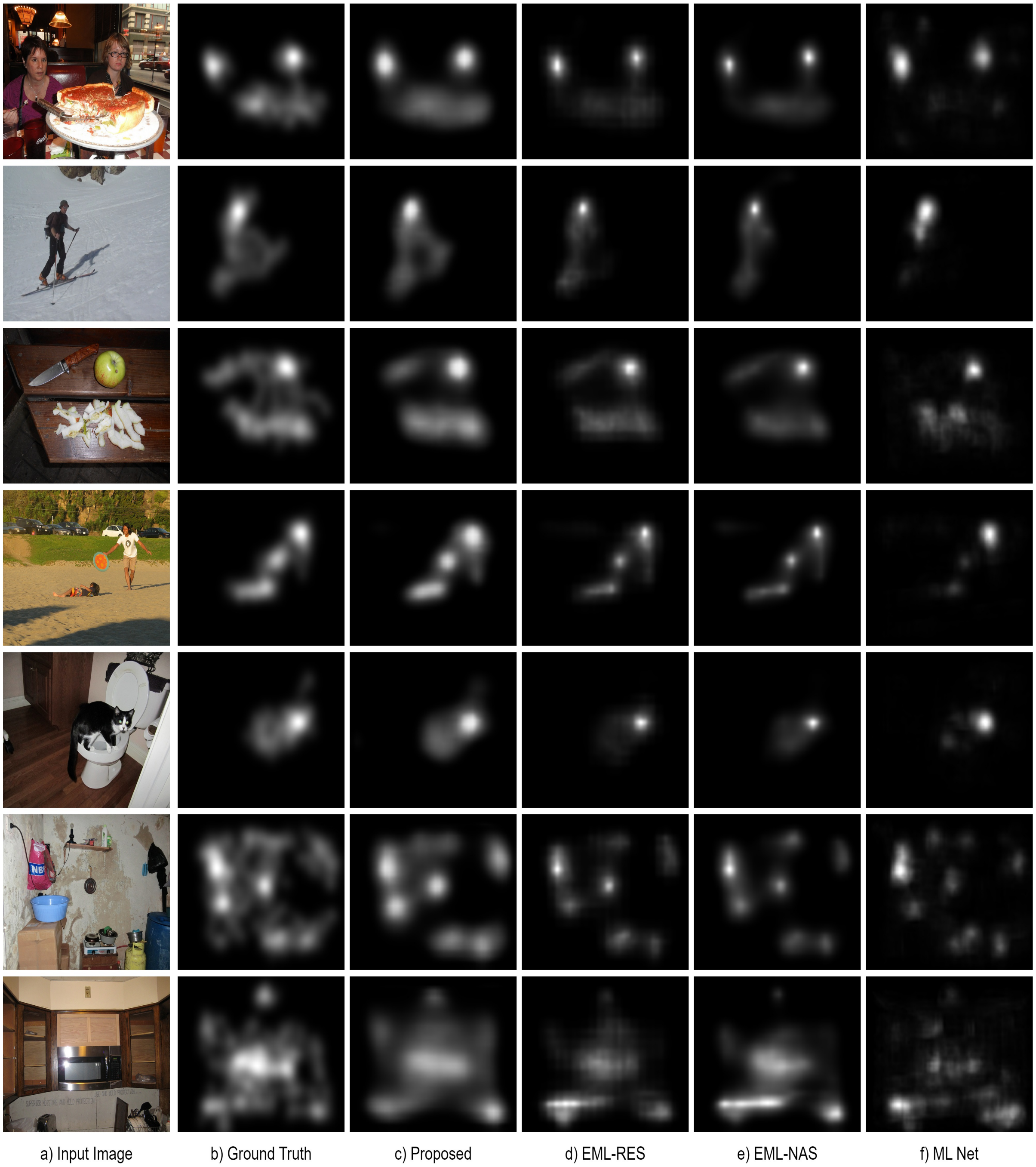

Figure 3. Qualitative Comparison of Fixation Maps: FixPNet Vs Competitor Methods

world data variations through the effective exploitation of prior knowledge. The inclusion of prior information aids in the capture of deep semantics. The improved prediction accuracy and low model complexity of the proposed method is highly suitable for solution deployment in low-power devices.

References

[1] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Frédo Durand. What do different evaluation metrics tell

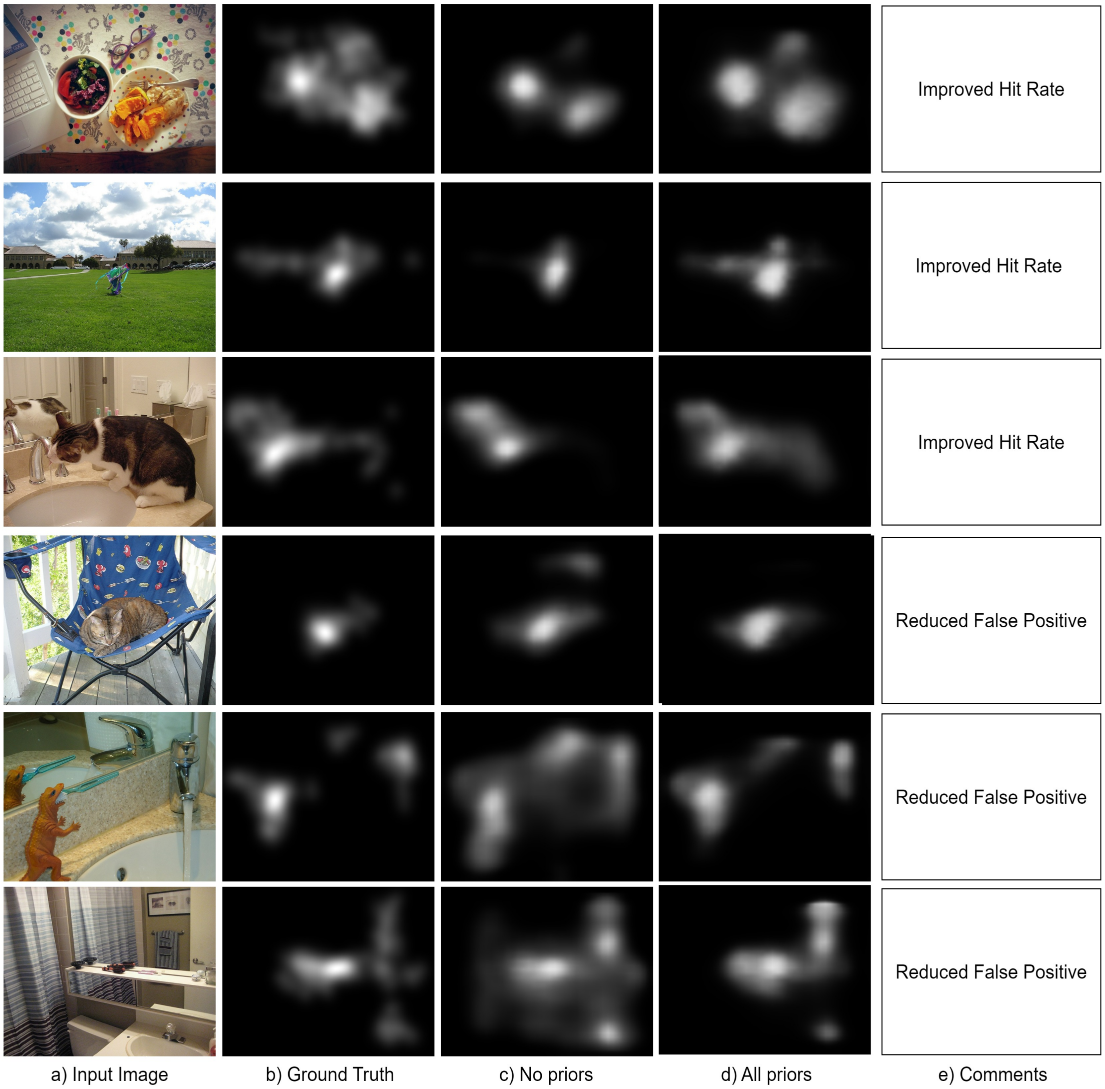

Figure 4. Examples Inputs and their Eye Fixation Maps

us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2019. 5

[2] Zhaohui Che, Ali Borji, Guangtao Zhai, Xiongkuo Min, Guodong Guo, and Patrick Le Callet. How is gaze influenced by image transformations? dataset and model. *IEEE Transactions on Image Processing*, 2019. 5

[3] 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

[4] 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

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

[6] 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

[7] 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

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

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

[10] 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

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

[12] Laurent Itti, Christof Koch, and Ernst Niebur. A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. Pattern Anal. Mach. Intell.*, 1998. 2

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

[14] 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, 4

[15] 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, 4

[16] Christof Koch and Shimon Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 1985. 2

[17] 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

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

[19] Min Seok Lee, WooSeok Shin, and Sung Won Han. TRACER: extreme attention guided salient object tracing network. *CoRR*, abs/2112.07380, 2021. 1, 3, 4

[20] 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

[21] Mingxing Tan and Quoc V. Le. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR*, abs/1905.11946, 2019. 3, 5

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

[23] 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

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

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