Emergence of Proto-Object Representations via Fixations in Low-Resolution

Chengyao Shen, Xun Huang and Qi Zhao

Abstract—One prominent feature of our visual system is that the fovea – the highest-resolution portion of the retina – only occupies two visual degrees, while the remaining portion of the retina (parafovea and periphery) are mainly in low-resolution. Therefore, before we make a saccadic eye movement, the potential fixation target is usually located in parafovea or periphery and is perceived in low-resolution. In this work, we present a computational framework based on convolutional neural network (CNN) to model this selective visual attention mechanism. By training the network with low-resolution inputs on potential fixation targets and non-salient locations, we find that proto-object representations emerge as a natural outcome for saliency prediction. These proto-object representations, which usually encode object gists and high-order statistics of a local region, demonstrate outstanding performance in predicting real eye fixation locations over other state-of-the-art saliency models. The component analysis also provides, good insight into the validity of our approaches in improving the performance of the model.

Index Terms—

1 Introduction

1.1 Motivation

Visual acuity of retina drops rapidly from central vision to peripheral vision. The fovea, which is the highest-resolution portion of the retina, only occupies two visual degree in the visual field[1], while the remaining portion of the retina (parafovea and periphery) are mainly in low-resolution. According to the relative visual acuity in human eye in Fig. 1(b), the highest visual acuity of retina drops by a factor of two at 2.5 visual degree and five at 10 visual degree [1], [2]. Hence, at one specific moment, our visual perception of a natural scene would only have high visual acuity in the center of the gaze while the remaining parts are sampled in low visual acuity (as illustrated in Fig. 1(a)). To remedy this information degeneration in input, our brain employs a strategy of selective visual attention to build up our visual perception. In our daily life, our eyes could efficiently select potential fixation targets and constantly make saccadic eye movements to construct a continuous high-resolution perception of our visual environment. These facts suggest that: (1) Targets at potential fixation locations (yellow dashed circle in Fig 1(a)) are often in low-resolution when they first enter visual system and trigger the saccadic eye movements. (2) Fixation locations are not selected in random, but according to some image statistics that could still be preserved in low-resolution.

C. Shen, X. Huang and Q. Zhao is with the Department of Electrical and Computer Engineering, National University of Singapore, Singapore

Corresponding author: Q. Zhao (email: eleqiz@nus.edu.sg).

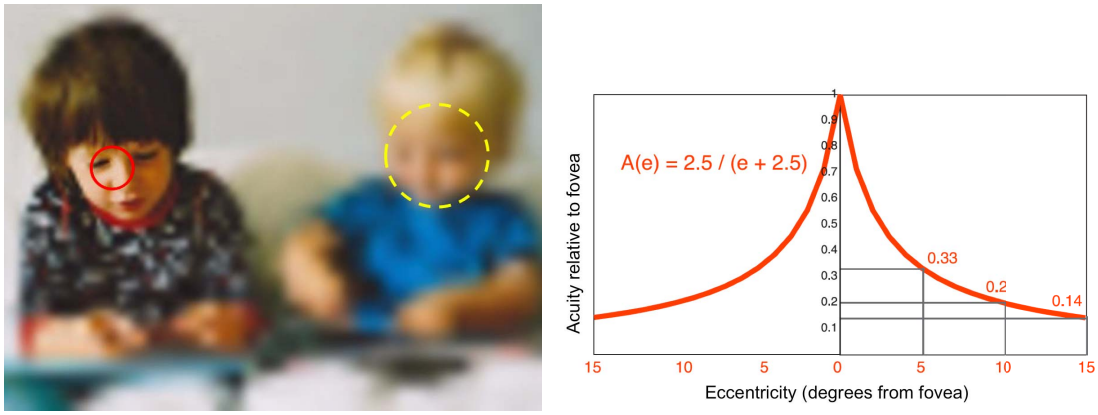
Manuscript received ; revised

1. approximately twice the width of the thumbnail at arm's length

1.2 Background

Classical views on human visual perception divide selective visual attention in two stages [3]: a "pre-attentive" stage that processes the visual information over the entire visual field in a fast and parallel way, and an "attentive" stage that is local, serial and associated with complex shape analysis and object recognition. According to Feature Integration Theory [4] (FIT), the "pre-attentive" stage extracts multi-scale low-level features (*e.g.* contrast, color, orientation, motion, etc.) in a parallel way and in the "attentive" stage, these features are bound together to an object conception. The core idea of FIT is that attention is driven by low-level features. Numerous computational models of attention are built upon this theory [5]–[8].

However, recent studies show that objects could be better fixation predictors than purely low-level features [9], especially when objects and strong low-level feature contrast are disjointed in location, or an object is of semantic meaning while its corresponding features are in low contrast. There are also theories and a growing amount of evidence showing that there is a "proto-object" representation at the pre-attentive stage of visual



(a) (b)

Fig. 1. Illustration of visual acuity falloff: (a) Illustration of a natural scene with visual acuity falloff. Red circle indicates the 2 visual degrees of fovea size at gaze position and yellow dashed circle indicate potential fixation location perceived at periphery. (b) Relative acuity of the human eye in eccentricity (degrees from fovea) [1].

perception and can be used to guide attention [10]–[13]. In definition [13], "proto-objects" can be seen as pre-attentive structures coherent in limited space and time. They can bind various low-level features over a small region of space and a short period of time and become "highest-level output of low-level vision". Unlike precise object recognition that happens after the deployment of attention and requires serial fovea processing (scrutinizing), proto-object is more like object gist [14] which approximates an object or a object cluster and can be computed rapidly in parallel over the entire visual field.

1.3 Our Approach

In this work, we present a computational framework based on convolutional neural network (CNN) to model the mechanism of fixation target selection in low visual acuity. By modelling saliency prediction as binary classification and training the network with low-resolution inputs that is close to the visual acuity at parafovea and periphery, we find that our model naturally learns out proto-object representations like object blobs, text patterns and human profiles. Evaluated on 4 datasets with 3 different evaluation metrics, this model demonstrates outstanding performance in predicting eye fixation locations over other state-of-the-art saliency models. Visualizations of the network also show interesting findings that are consistent with the hierarchy of early visual features, mid-level features and proto-object-like features in the "Coherence Theory" of proto-object in visual attention and visual search [13]. The main contributions of this work are summarized as follows:

1) We design for saliency prediction, a retina-inspired architecture that processes visual input with multiple low resolutions. It well simulates selective attention in humans where regions of interest that drive gaze are mostly at parafovea or periphery thus in low resolution.
2) We learn out meaningful proto-object representations from large-scale attention data under the proposed framework of fixation on low resolution.
3) We explore large-scale attention dataset as well as data augmentation methods on the dataset, and report prediction results with different data sizes for training a CNN-based saliency model.

2 Related Works

A commonly referred line of early models of visual attention were built upon the "Feature Integration Theory" (FIT) [4], and the conspicuity of a region or object is encoded in terms of early features such as color, orientation, depth and direction of motion [3]. These models mainly used a single layer of hand-crafted features such as multi-scale luminance contrast, color contrast and edge orientation [5]–[8]. In addition to the models using hand-crafted features, there are also models that use a single layer of features learned from natural image statistics. For these models, independent component analysis (ICA) and sparse coding are commonly used for feature learning. For example, SUN [15], ICL [16] and AIM [17] used features learned from ICA with links to information theory. Borji and Itti [18] built their model on features learned with sparse coding.

Despite the constant progress in saliency prediction based on low-level features, mounting evidence shows that the allocation of eye fixations do not only depend on low-level features but also on the structural organization of these low-level features into perceptual objects [11]–[13]. Psychophysics experiments also show that objects

can attract eye fixations better than simply low-level features [9].

To bridge the gap between low-level feature based saliency models and human behavior in consistently being attracted by object-level features, a number of recent models leveraged sophisticated object detectors and combined them with the original low-level feature framework [19]–[23]. However, these object detectors implicitly assume that the objects are already recognized before the saliency map generation, whereas pre-attentive selection is believed to act rapidly and not after the exact recognition of various objects [24]. Besides, considering the thousands of object categories existing in our daily life, simply adding detectors would make the saliency models intractable to implement.

Another line of models tackle this problem by extracting pre-segmented objects or proto-object [13] in the scene [11], [12], [24]–[26]. These models extract proto-object representations either by bio-inspired features [11], [24] or by computer vision techniques like hierarchical image segmentation [12], [26] and approximated ellipse [25]. The models that utilize bio-inspired features are closely related to our work. However, either the multi-scale features [24] or the "border ownership cells" and "grouping cells" [11] used in these models is hand-designed with no learning involved.

The recent development of deep learning models provides a coherent framework from efficient representations from low-level to object-level [27]. Deep learning models are generally multi-layer networks which can learn multi-level features from data. Instead of using pre-defined features or operations, deep learning models have the potential to learn task-related representations directly from data without any assumptions on the type of the features. Works that leverage the representations from deep learning to modeling human attention include Shen *et al.* [28], [29], Vig *et al.* [30], Kummerer *et al.* [31], and Liu *et al.* [32]. Shen *et al.* [28], [29] utilized a three-layer sparse coding network to learn a hierarchy of features on fixation regions and to predict saliency based on the learned features. Their network is learnt in a layerwise unsupervised manner while ours is with mutli-scale back propagation. Vig *et al.* [30] trained integration weights for an ensemble of randomly initialized networks and did large-scale optimal network search. Kummerer *et al.* [31] leveraged a sophisticated convolutional neural network pre-trained on precise object classification, while we adopt a different model structure whose features/parameters are learned directly from fixations, for saliency prediction. Liu *et al.* [32] used shared-weight CNN in multi-resolution, yet with a CNN structure quite different from ours (e.g., in the final layer and the scale). Their saliency prediction is based on a grid sampling while ours is on full image convolution, which is computationally more efficient.

3 METHODS

3.1 Low-Resolution Inputs

We model the visual acuity of parafovea and periphery with multi-scale low-resolution inputs. We extract low-resolution image patches in multiple visual acuity from potential fixation targets and non-target locations on the image according to the "sunflower" model of retina [2], [33], [34]. The "sunflower" model is a stacked model for the retinal receptive fields that tackles the scale-space property of retina sampling. To determine the locations of potential fixation targets and non-targets, the ground truth attentional map is first convolved with a gaussian mask whose standard deviation is 1 visual degree (around 24 pixels), the locations are then selected as follows:

1) **Potential fixation target locations:** the ground truth attentional map is first convolved with a gaussian mask whose standard deviation is 1 visual degree (around 24 pixels). The top five local maxima in the blurred ground truth maps are then used as potential fixation target locations. If local maxima are less than one third of the global maxima, it would be abandoned and in this case there would be less than 5 potential fixation targets on that image
2) **Non-target locations:** in each blurred ground-truth map, five locations randomly sampled from the pixels with saliency values less than the mean of the whole map are selected as non-target locations.

We then extract multi-scale image patches centered at locations of potential fixation target or non-target. Specifically, we extract patches from training images with an increasing size of $\sqrt{2}$ and then downsample them to a same size to yield a relative visual acuity of $\{0.5, 0.25\sqrt{2}, 0.25, 0.125\sqrt{2}, 0.125\}$. For example, in the 3-layer model whose training patches are in the size of 36×36, we extract patches centered at the location of a potential fixation target with a size of $\{72 \times 72, 102 \times 102, 144 \times 144, 204 \times 204, 288 \times 288\}$ and downsample them to 36×36.

Data Augmentation: To avoid the problem of overfitting in CNN training, we adopt data augmentation, adding specific image transformations to input images and ground truth attention maps simultaneously before the image patch extraction. The image transformations we use include a horizontal image flip and a rotation transformation with a random uniform distribution between -15° and 15°, which increases the variability

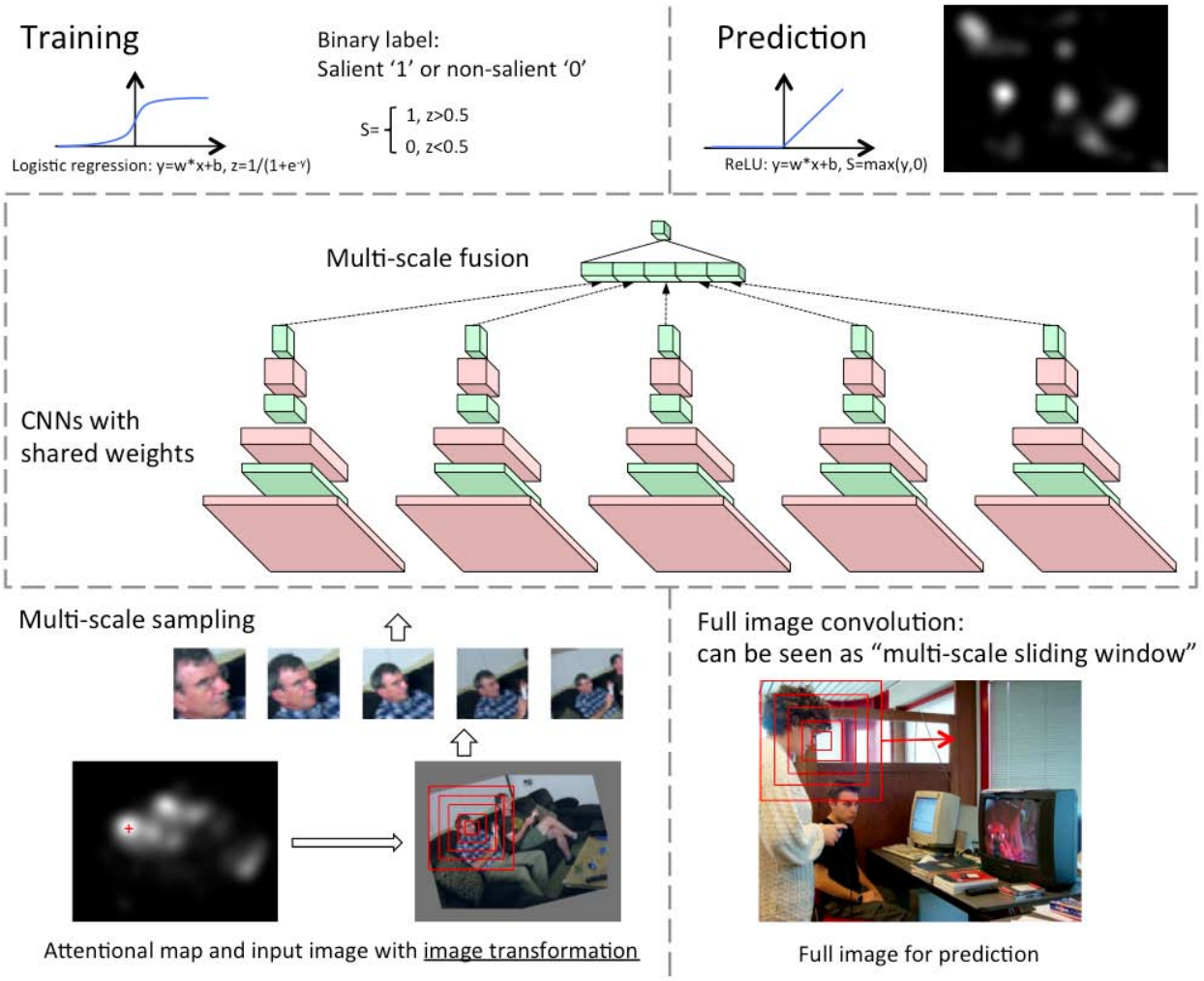


Fig. 2. Network structure of the 3-layer model. Red blocks indicate the responses of convolution followed by a rectified linear unit, green blocks indicate the responses of pooling operation. In the training stage, multi-scale salient and non-salient patches are input into the network and the network is trained as a binary classification with logistic regression. In saliency prediction stage, a full image is fed into the network and a saliency map is generated with trained parameters.

of inputs while still keeping the inputs semantically unchanged.

3.2 The Model

Our model leverages a weight-sharing multi-scale convolutional neural network (CNN) scheme (as illustrated in Fig. 2): CNNs with shared weights are fed with low resolution image patches at different visual acuity and their output feature maps are concatenated at the final stage and then linearly integrated into a final saliency map. A logistic regression is used as the final step to model the training process as a binary classification on potential fixation targets and non-targets.

Three CNN structures are used for each single scale:

- 2-layer model: C(5,64)-MP(2)-C(5,512)-MP(2)
- 3-layer model: C(5,64)-MP(2)-C(5,128)-MP(2)-C(5,512)-MP(2)
- 4-layer model: C(5,64)-MP(2)-C(5,128)-MP(2)-C(5,256)-MP(2)-C(5,512)-MP(2)

where C(f,n) indicates n convolution kernels in the size of $f \times f$, MP(f) indicates non-overlap max pooling in $f \times f$.

Training: In the training stage, the input patch sizes for these three structures are 16×16, 36×36, 76×76 respectively. These sizes are set to ensure that the final output for each single scale is in the size of 1×1. In the training stage, we use convolutions without padding to avoid the edge effect that will influence the generalization ability of the model in saliency prediction. We have also explored different variations on the structure including adding spatial normalization and using all-layer integration. These variations are bio-inspired, but do not yield a better performance of the model in saliency prediction.

Saliency Prediction: For salient prediction, we feed

a full image into the weight-sharing multi-scale CNN and get the output of the network as a saliency map. Different from the training stages where we use logistic regression for binary classification, in saliency prediction we remove the final sigmoid function in logistic regression and use a rectified linear integration instead as the output:

$$\mathbf{S} = \max(w * \mathbf{x} + b, 0) \quad (1)$$

where $\mathbf{x}$ is the network responses after multi-scale concatenation, w and b are the integration weights and bias respectively and $\mathbf{S}$ is the generated saliency map. This operation helps to avoid information loss in sigmoid function and to remove the unnecessary details on non-target regions. In saliency prediction, we use padded convolution with a padding size of 2 to ensure that the output saliency maps have the same aspect ratio with the input images.

4 Experiments and Results

This section reports experimental results and analyses to validate the representations learned out in the network and the performance of our model on eye fixation prediction. We train our model on mulit-scale low-resolution patches extracted from the SALICON dataset [35] and validate it on MIT1003 [19], OSIE [36], FIFA [20] and NUSEF [37]. The saliency maps computed from these 4 datasets are then evaluated quantitatively with other state-of-the-art eye fixation prediction algorithms. Visualizations of features at each layer and component analysis of the proposed model are also provided.

4.1 Datasets and Training

The Saliency in Context (SALICON) dataset [35] is a recently published dataset containing 10,000 images from the Microsoft Common Object in Context (COCO) [38] dataset. In the SALICON dataset, large-scale mouse movement data from human free-viewing an image is recorded through Amazon Mechanic Turk (AMT). With a new psychophysics method, the mouse trajectories of subjects can indicate where people look in the images. It is demonstrated in their experiments that the mouse maps generated from this mouse movement data are highly consistent with eye fixation data and can be used as ground-truth for training and evaluating saliency models.

During training, the 10,000 images in the SALICON dataset are used as training set. 100 images from OSIE dataset [36] are used as validation set to monitor the progress of training. We train the models with 200 epochs, with potential fixation targets and non-targets from 2000 images in an epoch. The final parameters used for saliency prediction is the one with the lowest validation objective in 200 epochs. We leverage the MatConvNet toolbox [39] to implement our models. With GPU acceleration, it normally takes half a day or one day to train a model.

For saliency prediction, we test our models on 4 standard eye tracking datasets. The MIT1003 dataset [19] contains 1003 landscape and portrait images. The OSIE dataset [36] contains 700 images of natural scenes and aesthetic photographs. The eye movement data in both datasets were collected from 15 observers during free-viewing. The FIFA [20] dataset consists of 200 images free-viewed by 8 observers. The NUSEF [37] dataset contains 758 images with affective objects, where image is viewed by 25 subjects on average.

4.2 Evaluation Metrics

The evaluation metrics we use include shuffled Area Under Curve (sAUC), linear Correlation Coefficient (CC) and Normalized Scanpath Saliency (NSS) [23], [40].

AUC is the most widely used score for saliency model evaluation. In the computation of AUC, the estimated saliency map is used as a binary classifier to separate the positive samples (human fixations) from the negatives (random points). By varying the threshold on the saliency map, a Receiver Operating Characteristics (ROC) curve can then be plotted as the true positive rate vs. false negative rate. AUC is then calculated as the area under this curve. However, AUC can be easily influenced by center-bias in the human ground truth data. **sAUC** (shuffled AUC) is the same as AUC except using fixations of other images in the same dataset as negatives and is able to eliminate the effect of center-bias.

CC measures the linear correlations between the estimated saliency map and the ground truth fixation map. The closer CC to 1, the better the performance of the saliency algorithm.

NSS measures the average of the response values at fixation locations along the scanpath in the normalized saliency map. The larger the NSS score, the more corresponding between predictions and ground truths.

All these metrics have their own advantages and limitations and a model that performs well should yield high scores in all these metrics. In all our comparison and analysis we did not include explicit center bias to the saliency map to ensure a consistent and fair comparison across models.

4.3 Model Performance

We compared the proposed models with the state-of-the-art ones (with codes available). These models include BMS [26], eDN [30], Judd [19], AWS [8], AIM [17], GBVS [41], LG [18], SigSal [42], and Itti [5].

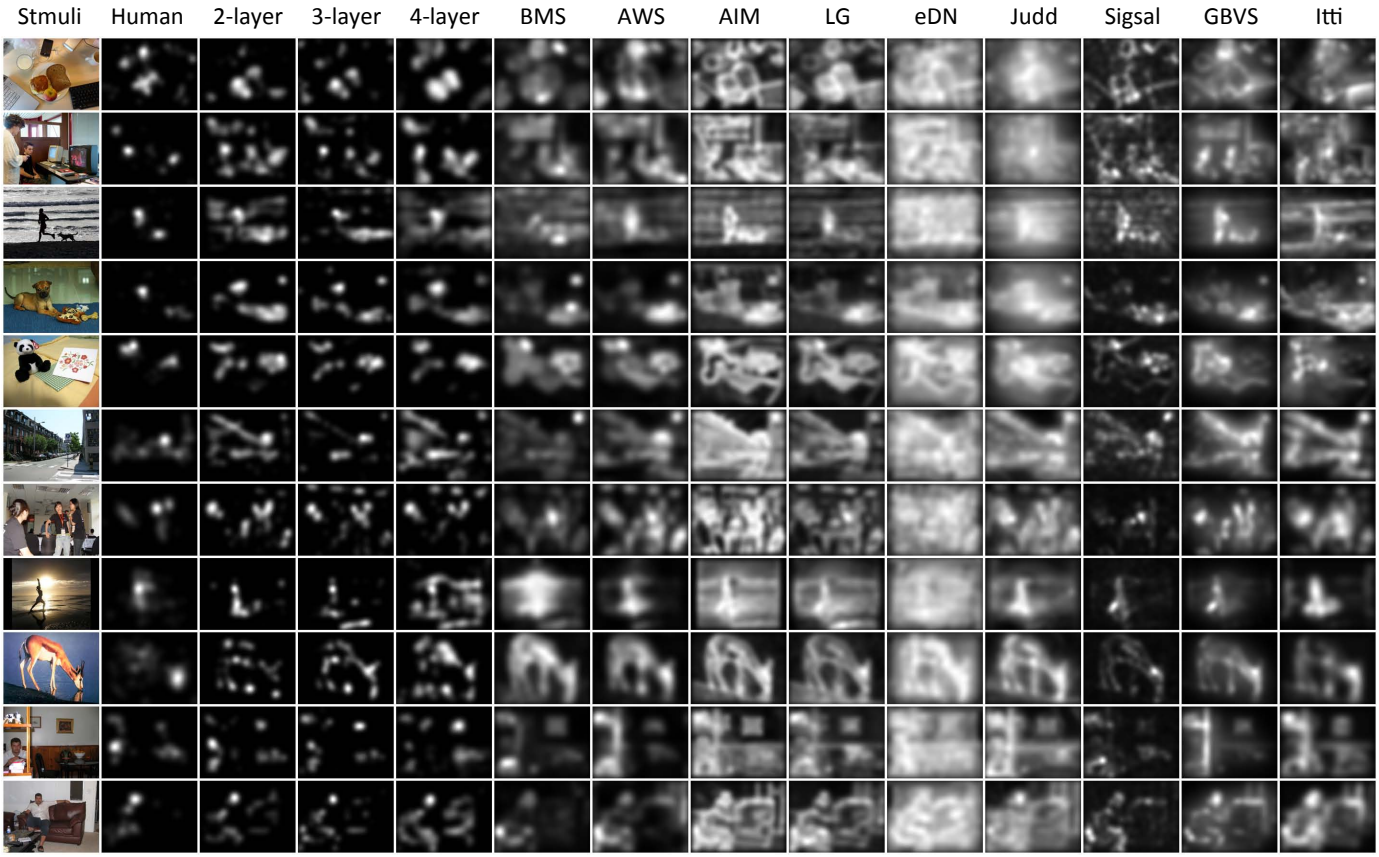


Fig. 3. Qualitative comparison of our models with human ground truth and other state-of-the-art algotrithms on different images from MIT1003, OSIE, NUSEF and FIFA datasets. The models are in general able to detect various objects in natural scene images.

	OSIE			MIT1003			NUSEF			FIFA		
	sAUC	CC	NSS	sAUC	CC	NSS	sAUC	CC	NSS	sAUC	CC	NSS
2-layer	0.783	0.567	2.010	0.694	0.533	1.438	0.646	0.610	1.426	0.790	0.555	2.172
3-layer	**0.817**	**0.606**	**2.236**	**0.718**	**0.577**	**1.588**	**0.655**	**0.641**	**1.518**	**0.816**	**0.609**	**2.387**
4-layer	0.800	0.534	1.905	0.705	0.486	1.316	0.642	0.520	1.220	0.782	0.461	1.845
BMS	0.764	0.468	1.478	0.687	0.491	1.234	0.632	0.546	1.203	0.756	0.422	1.359
AWS	0.764	0.453	1.452	0.686	0.445	1.107	0.628	0.492	1.096	0.745	0.370	1.216
eDN	0.730	0.375	1.129	0.675	0.458	1.063	0.621	0.502	1.057	0.736	0.362	1.115
Judd	0.667	0.404	1.253	0.665	0.456	1.095	0.620	0.512	1.116	0.761	0.405	1.308
LG	0.753	0.417	1.306	0.678	0.427	1.033	0.618	0.472	1.024	0.737	0.364	1.123
SigSal	0.732	0.423	1.319	0.666	0.465	1.085	0.614	0.495	1.094	0.747	0.402	1.268
AIM	0.754	0.413	1.254	0.680	0.469	1.082	0.629	0.491	1.054	0.760	0.392	1.210
GBVS	0.697	0.431	1.359	0.643	0.502	1.254	0.591	0.559	1.204	0.716	0.425	1.352
ITTI	0.644	0.294	0.851	0.645	0.468	1.127	0.577	0.305	0.642	0.690	0.384	1.165

TABLE 1
Performance of different models on MIT1003, OSIE, NUSEF and FIFA datasets. The highest scores are in bold.

From the quantitative results in Table 1, it can be seen that all the three proposed models consistently outperform other algorithms, and the 3-layer model performs the best.

From qualitative comparison shown in Fig. 3, it can be observed that visually our predicted saliency maps, especially the maps from the 3-layer model, are more similar to the ground truth than the maps from the other models. The proposed models (the 2-layer, 3-layer and 4-layer models) are in general able to detect various objects in natural images, while the other models would also have strong responses on various low-level features. Of the three proposed models, our 3-layer model performs better in cluttered background (see the 3rd row and the 6th row of Fig. 3). Yet for the saliency maps from other algorithms, we can observe that the false responses at background or object edges are usually strong, which are quite different from the human ground truth.

4.4 Feature Visualization

To further characterize the networks learned on low-resolution inputs, we visualize the features in the 3-layer model to see what has been learned out.

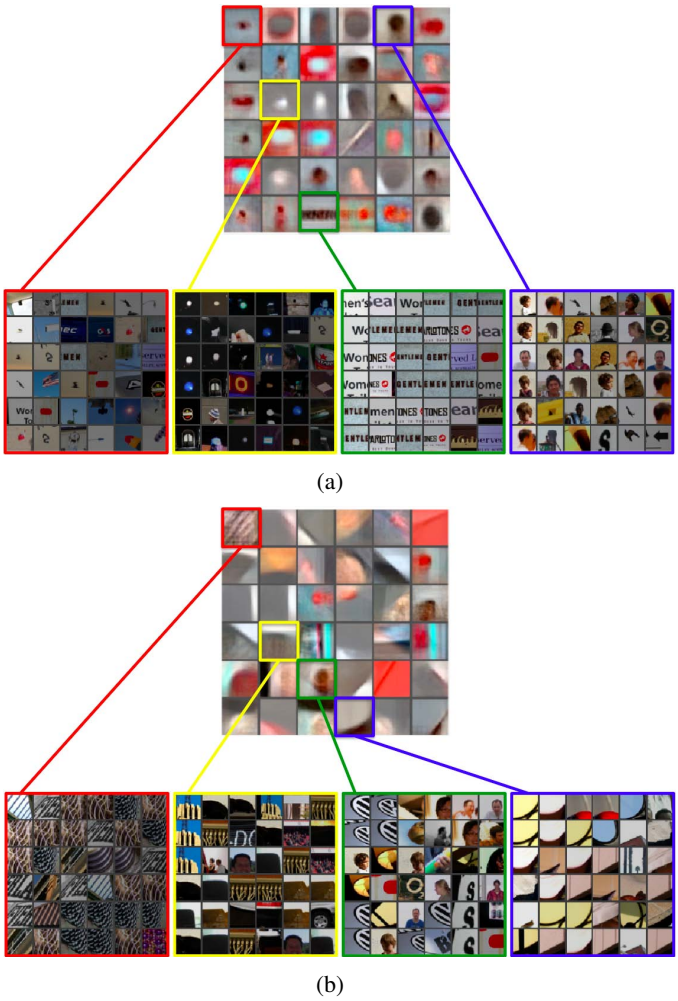
For features in layer 1, they are visualized directly with their weights in the convolutional layer since the layer 1 is directly connected to the input space. However, for features in higher layers, the direct visualization of feature weights would not explicitly reveal the input patterns represented by the features. Therefore, we visualize the features in higher layers by the averages of top responsive inputs as described in [29]. More specifically, we traverse a large number of images in the training set and generate the responses of each convolution layer. Top responsive neurons in one specific layer are then selected and their corresponding effective input are cropped out and averaged for feature visualization.

Low-level features: The visualization of features in layer 1 are illustrated in Fig. 4(a). They are mostly luminance, color and edge like features with different spatial frequency, which are typical low-level features used in saliency literature and correspond to neural findings well.

Mid-level features: For features in layer 2, we visualize them with the average of top 64 responsive inputs. From Fig. 4(b), it can be observed that apart from long edges, mid-level features like curvatures and junctions are also learned out.

Proto-object-like features: For features in layer 3, we visualize them with the averages of top 100 responsive inputs. From Fig. 4(c), many object blob-like representations are learned out.

In the 3-layer model, the layer 3 extract 512×5 feature maps from each images and the linear integration layer integrates them with weights on each single feature map. Hence, these features directly contribute to saliency and their contribution can be measured by the weights of the linear combination layer. To gain more insights into the contribution of these features to saliency prediction, we select the features in layer 3 with top 36 positive weights and top 36 negative weights, which represent the most target-like representations and most non-target like representations. Fig. 5(a) shows that the target-like representations are mostly object center or proto-object-like representations. The 4 features generally represent object blobs in light background, object blobs in dark background, text-like patterns and head-like patterns. These features are not selective to one specific category and demonstrate an explicit proto-object representation. In Fig. 5(b), we observe that the non-target respresentations are mostly textures and edges which are likely to appear at background or object contours.



(a)

(b)

Fig. 5. (a) Visualization of features in layer 3 with top 36 positive weights and expanded illustration of 4 typical features for potential fixation targets. (b) Visualization of features in layer 3 with top 36 negative weights and expanded illustration of 4 typical features for non-targets

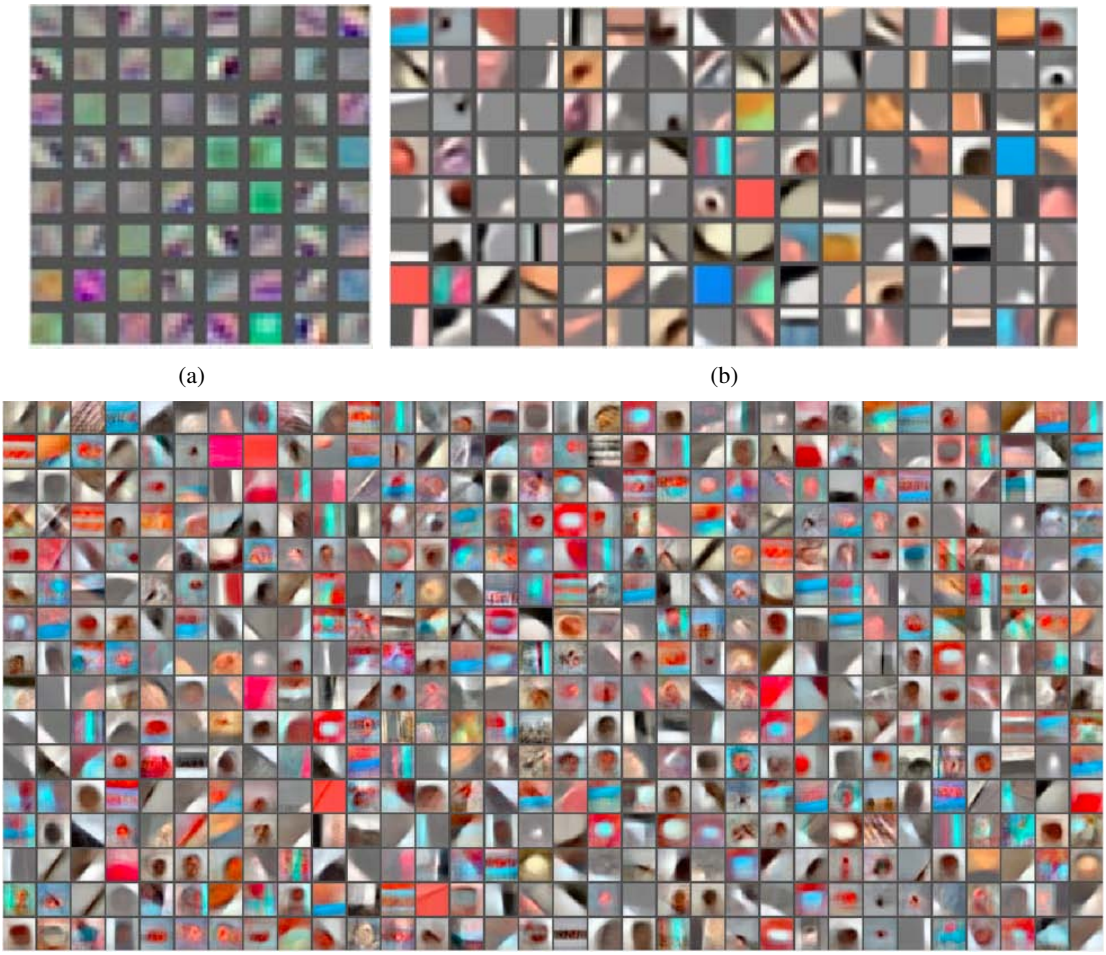


Fig. 4. Visualization of (1) layer 1 features, and (b) layer 2 features, and (c) layer 3 features

In a previous work [43], [44] that aims at learning regularities in eye movement data with a single layer model, it is found that center-surround patterns would emerge as optimal predictors for potential fixation targets. The discovery of their model is interesting and the results of our models are consistent with their results. From Fig. 4(c), it can be observed that many proto-object representations would display center-surround patterns. However, with the expanded view of each feature in Fig. 5(a), it can be seen that our proto-object representations could encode more high-order statistics of a local region. Compared with one single layer of center-surround filters, these proto-object filters are more selective to certain type of potential fixation targets, which more or less results in the better performance of our models in predicting eye fixations.

4.5 Component Analysis

4.5.1 Scale and Multi-Scale Fusion

To gain more insights into the contribution of each single scale in saliency prediction. We train 4 models with training data only from 4 single scale and measure these models' performance on the OSIE and MIT1003 datasets. From Table 2 and Table 3, we observe that the scales 0.5 and 0.25 have the largest contributions to saliency prediction performance. The performance on the finest scale drops a bit and the coarsest scale has the worst performance. The results indicate the scale with middle resolution contributes most to eye fixation prediction, while the scale in very fine or coarse resolutions contribute less to saliency.

Scale	sAUC	CC	NSS
1	0.790	0.550	2.019
0.5	0.798	0.571	2.119
0.25	0.802	0.576	2.101
0.125	0.760	0.491	1.621

TABLE 2
Single scale comparison on OSIE dataset

We also visualize the saliency maps generated from single scale models in Fig. 6. We could see that coarse

Scale	sAUC	CC	NSS
1	0.698	0.473	1.332
0.5	0.702	0.496	1.392
0.25	0.702	0.491	1.349
0.125	0.671	0.400	1.056

TABLE 3
Single scale comparison on MIT1003 dataset

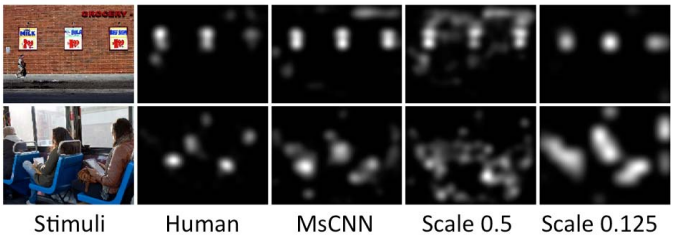


Fig. 6. Visualization of results generated from single scale models. Human ground truth and results of multi scale model are also illustrated for comparision.

scale is in general good at generating reasonable saliency maps while fine scale may introduce noise. However, the coarse scale sometimes neglect small but salient regions in a scene. Hence, the application of multi-scale leverage the advantage from both scales.

To further verify this, we expanded the scale range of our reported model in both fine and coarse directions in a step of $\sqrt{2}$ time. From Table 4 and Table 5, we can find the performance of adding more scale on finer side ('1-0.125 concat') does not turn better while the performance of adding one more scale in coarse side ('0.5-0.0625 concat') lead to worse scores on NSS and CC. In this table, we also compare the performance of different fusion methods and we found that the performance of cross scale pooling is a little bit worse that that of multi-scale concatenation. The results are consistent across different datasets.

Scale Range	sAUC	CC	NSS
0.5-0.125 concat	0.817	0.606	2.236
0.5-0.125 pool	0.812	0.599	2.187
1-0.125 concat	0.818	0.614	2.286
0.5-0.0625 concat	0.820	0.583	2.069

TABLE 4
Multi-scale comparison on OSIE dataset

4.5.2 Effect of Image Transformations

In this subsection, we analyze how different image transformations in online data augmentation would affect performance of our model. As shown in Fig. 7, we find that both the horizontal flip and the affine transformations would improve the model performance in a large margin.

Scale Range	sAUC	CC	NSS
0.5-0.125 concat	0.718	0.577	1.588
0.5-0.125 pool	0.710	0.557	1.542
1-0.125 concat	0.715	0.561	1.562
0.5-0.0625 concat	0.713	0.512	1.396

TABLE 5
Multi-scale comparison on MIT1003 dataset

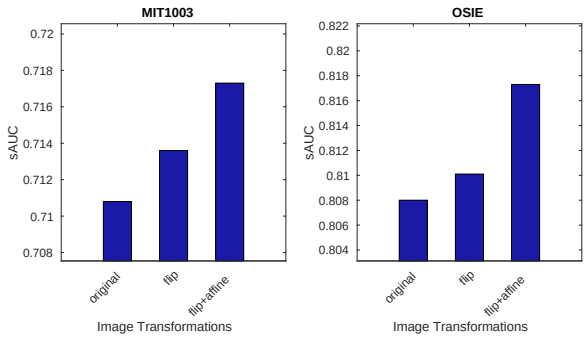


Fig. 7. The relationship between image transformations and sAUC scores.

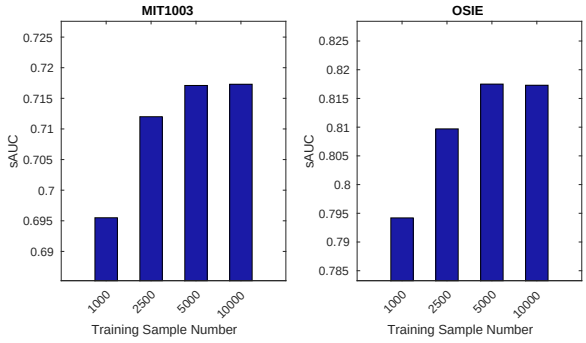


Fig. 8. The relationship between training sample number and sAUC scores.

4.5.3 Training Set Size

To demonstrate the effect of number of training samples on performance, we train four networks using 1,000, 2,500, 5,000 and 10,000 training images respectively. Other parameter settings of the networks are the same. We evaluate the four networks on four datasets and summarize the results in Fig. 8. As we expected, the performance on all datasets increases with the number of training samples. However, the score improves only slightly from 5,000 training images to 10,000 training images, indicating a saturation in sample size. Typical large eye tracking datasets contain around 1000 images.

According to our results, there is a big improvement from using 1,000 training images to 10,000 images, indicating the effectiveness of training with large scale attentional data.

5 Conclusion

This paper presents a new computational model to effectively learn features from potential fixation targets in low resolution. A saliency model based on the multi-scale low-resolution CNN framework is further proposed and demonstrated to be competitive and promising in predicting where people look at. Results demonstrate that, by training to differentiate potential fixation targets and non-targets in low resolution, proto-object representations can be learned in a multi-layer architecture similar to conceptual models of visual attention in the literature [13], [45].

Acknowledgments

This research was supported by he Singapore Ministry of Education Academic Research Fund Tier 2 (No.R-263-000-B32-112) and the Defense Innovative Research Programme (No. 9014100596)

References

[1] T. Judd, F. Durand, and A. Torralba, "Fixations on low-resolution images," *Journal of Vision*, vol. 11, no. 4, p. 14, 2011.

[2] T. Lindeberg and L. Florack, "Foveal scale-space and the linear increase of receptive field size as a function of eccentricity," 1994.

[3] C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," *Hum. Neurobiol.*, vol. 4, no. 4, pp. 219–27, 1985.

[4] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[5] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[6] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in neural information processing systems*, pp. 545–552, 2006.

[7] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision research*, vol. 49, no. 10, pp. 1295–1306, 2009.

[8] A. Garcia-Diaz, X. R. Fdez-Vidal, X. M. Pardo, and R. Dosil, "Saliency from hierarchical adaptation through decorrelation and variance normalization," *Image and Vision Computing*, vol. 30, no. 1, pp. 51–64, 2012.

[9] W. Einhäuser, M. Spain, and P. Perona, "Objects predict fixations better than early saliency," *Journal of Vision*, vol. 8, no. 14, 2008.

[10] R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual review of neuroscience*, vol. 18, no. 1, pp. 193–222, 1995.

[11] A. F. Russell, S. Mihalaş, R. von der Heydt, E. Niebur, and R. Etienne-Cummings, "A model of proto-object based saliency," *Vision research*, vol. 94, pp. 1–15, 2014.

[12] V. Yanulevskaya, J. Uijlings, J.-M. Geusebroek, N. Sebe, and A. Smeulders, "A proto-object-based computational model for visual saliency," *Journal of vision*, vol. 13, no. 13, p. 27, 2013.

[13] R. A. Rensink, "Seeing, sensing, and scrutinizing," *Vision research*, vol. 40, no. 10, pp. 1469–1487, 2000.

[14] J. A. Martins, J. Rodrigues, and J. du Buf, "Local object gist: meaningful shapes and spatial layout at a very early stage of visual processing," *GESTALT THEORY*, vol. 34, no. 3/4, 2012.

[15] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell, "Sun: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, 2008.

[16] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," *Advances in neural information processing systems*, vol. 21, pp. 681–688, 2008.

[17] N. Bruce and J. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, 2009.

[18] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 478–485, IEEE, 2012.

[19] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Computer Vision, 2009 IEEE 12th International Conference on*, pp. 2106–2113, IEEE, 2009.

[20] M. Cerf, J. Harel, W. Einhäuser, and C. Koch, "Predicting human gaze using low-level saliency combined with face detection," *Advances in neural information processing systems*, vol. 20, 2008.

[21] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *Journal of Vision*, vol. 11, no. 3, 2011.

[22] Q. Zhao and C. Koch, "Learning visual saliency," in *Information Sciences and Systems (CISS), 2011 45th Annual Conference on*, pp. 1–6, IEEE, 2011.

[23] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pp. 438–445, IEEE, 2012.

[24] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[25] M. Wischnewski, A. Belardinelli, W. X. Schneider, and J. J. Steil, "Where to look next? combining static and dynamic proto-objects in a tva-based model of visual attention," *Cognitive computation*, vol. 2, no. 4, pp. 326–343, 2010.

[26] J. Zhang and S. Sclaroff, "Saliency detection: a boolean map approach," in *Computer Vision (ICCV), 2013 IEEE International Conference on*, pp. 153–160, IEEE, 2013.

[27] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[28] C. Shen, M. Song, and Q. Zhao, "Learning high-level concepts by training a deep network on eye fixations," in *NIPS Deep Learning and Unsupervised Feature Learning Workshop*, vol. 2, 2012.

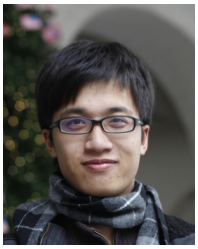
[29] C. Shen and Q. Zhao, "Learning to predict eye fixations for semantic contents using multi-layer sparse network," *Neurocomputing*, vol. 138, pp. 61–68, 2014.

[30] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," 2014.

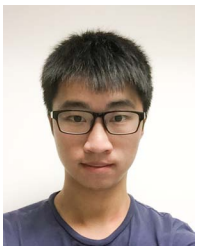
[31] M. Kümmerer, L. Theis, and M. Bethge, "Deep gaze i: Boosting saliency prediction with feature maps trained on imagenet," *arXiv preprint arXiv:1411.1045*, 2014.

[32] N. Liu, J. Han, D. Zhang, S. Wen, and T. Liu, "Predicting eye fixations using convolutional neural networks," in *Proceedings*

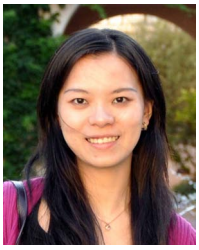
of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 362–370, 2015.
[33] J. Koenderink and A. Van Doorn, "Visual detection of spatial contrast; influence of location in the visual field, target extent and illuminance level," *Biological Cybernetics*, vol. 30, no. 3, pp. 157–167, 1978.
[34] B. M. ter Haar Romeny, "A scale-space model for the retinal sampling," *Front-End Vision and Multi-Scale Image Analysis: Multi-Scale Computer Vision Theory and Applications, written in Mathematics*, pp. 167–177, 2003.
[35] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2015.
[36] J. Xu, M. Jiang, S. Wang, M. S. Kankanhalli, and Q. Zhao, "Predicting human gaze beyond pixels," *Journal of Vision*, vol. 14, no. 1, pp. 1–20, 2014.
[37] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua, "An eye fixation database for saliency detection in images," in *Computer Vision–ECCV 2010*, pp. 30–43, Springer, 2010.
[38] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft coco: Common objects in context," in *Computer Vision–ECCV 2014*, pp. 740–755, Springer, 2014.
[39] A. Vedaldi and K. Lenc, "Matconvnet – convolutional neural networks for matlab," *CoRR*, vol. abs/1412.4564, 2014.
[40] Z. Bylinskii, T. Judd, F. Durand, A. Oliva, and A. Torralba, "Mit saliency benchmark." http://saliency.mit.edu/.
[41] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in neural information processing systems*, vol. 19, p. 545, 2007.
[42] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 1, pp. 194–201, 2012.
[43] W. Kienzle, F. A. Wichmann, M. O. Franz, and B. Schölkopf, "A nonparametric approach to bottom-up visual saliency," in *Advances in Neural Information Processing Systems*, pp. 689–696, 2007.
[44] W. Kienzle, M. O. Franz, B. Schölkopf, and F. A. Wichmann, "Center-surround patterns emerge as optimal predictors for human saccade targets," *Journal of Vision*, vol. 9, no. 5, p. 7, 2009.
[45] J. Duncan and G. Humphreys, "Beyond the search surface: Visual search and attentional engagement.," 1992.



Chengyao Shen received the BS degree in microelectronics from Shanghai Jiaotong University, China, in 2010. He is currently a Ph.D. candidate in the Visual Information Processing Lab, National University of Singapore. His research interests included computer vision, machine learning and natural image statistics.



Xun Huang is currently a visiting scholar at National University of Singapore. He is an undergraduate student at Beihang University and is expected to receive B.S degree in Computer Science in 2016. He has broad research interests in deep learning, computer vision and cognitive science.



Qi Zhao is an assistant professor in the Electrical and Computer Engineering Department at National University of Singapore (NUS) and the principal investigator at the Visual Information Processing Lab, working on computational vision and cognitive neuroscience. She also holds an appointment in the Ophthalmology Department and the Interactive and Digital Media Institute at NUS. She received the M.Sc. and Ph.D. degrees in computer engineering from the University of California, Santa Cruz, in 2007 and 2009 respectively. Prior to joining NUS, she was a postdoctoral researcher in the Computation & Neural Systems, and Division of Biology at the California Institute of Technology from 2009 to 2011. Her main research interests include computational vision, machine learning, computational cognition, and neuroscience. She has published more than 30 journal and conference papers in top computer vision, cognitive neuroscience, and machine learning venues, and is editing a book with Springer, titled "Computational and Cognitive Neuroscience of Vision", that provides a systematic and comprehensive overview of vision from various perspectives, ranging from neuroscience to cognition, and from computational principles to engineering developments. She is a member of the IEEE.