An Integrated Model for Effective Saliency Prediction

Xiaoshuai Sun,[1,2] **Zi Huang,**[1] **Hongzhi Yin,**[1] **Heng Tao Shen**[1,3]
[1]The University of Queensland, Brisbane 4067, Australia. [2]Harbin Institute of Tehcnology, Heilongjiang 150001, China.
[3]University of Electronic Science and Technology of China, Chengdu 611731, China.
xiaoshuaisun.hit@gmail.com, huang@itee.uq.edu.au, h.yin1@uq.edu.au, shenht@itee.uq.edu.au

Abstract

In this paper, we proposed an integrated model of both semantic-aware and contrast-aware saliency (**SCA**) combining both bottom-up and top-down cues for effective eye fixation prediction. The proposed **SCA** model contains two pathways. The first pathway is a deep neural network customized for semantic-aware saliency, which aims to capture the semantic information in images, especially for the presence of meaningful objects and object parts. The second pathway is based on on-line feature learning and information maximization, which learns an adaptive representation for the input and discovers the high contrast salient patterns within the image context. The two pathways characterize both long-term and short-term attention cues and are integrated using maxima normalization. Experimental results on artificial images and several benchmark dataset demonstrate the superior performance and better plausibility of the proposed model over both classic approaches and recent deep models.

Introduction

The last two decades have witnessed enormous development in the field of computational visual attention modeling and saliency detection (Borji and Itti 2013). Various models, datasets, and evaluation metrics are proposed to help machines better understand and predict human viewing behavior. By producing a 2D or 3D saliency map that predict where human look, a computational model can be applied to many low-level computer vision applications, e.g. detecting abnormal pattern(Itti, Koch, and Niebur 1998), segmenting proto objects(Hou and Zhang 2007), generating object proposals (Alexe, Deselaers, and Ferrari 2012) etc. The concept of "saliency" were investigated not only in early vision modeling but also in many engineering applications such as image compression (Itti 2004), object recognition (Salah, Alpaydin, and Akarun 2002) and tracking (Frintrop 2010), robot navigation (Siagian and Itti 2009), design and advertising (Rosenholtz, Dorai, and Freeman 2011) etc.

In the early stage, the research works are mostly motivated by biological priors (Koch and Ullman 1987; Itti, Koch, and Niebur 1998), statistical assumptions(Gao, Mahadevan, and Vasconcelos 2008) and information theory(Bruce and Tsotsos 2006; Hou and Zhang 2008). Some papers also draw their inspirations by analyzing the statistical property of ground-truth eye-tracking data (Sun et al. 2014; Leboran et al. 2016).

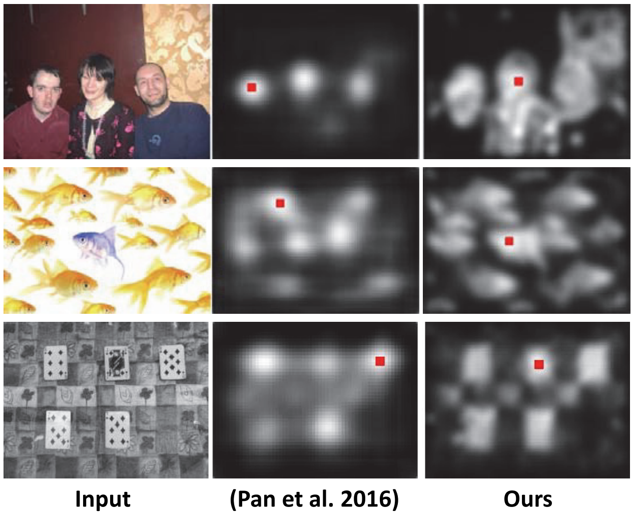

Figure 1: Illustration of our model. For each row, we show the input image (left), the saliency prediction result of (Pan et al. 2016) (middle) and our proposed **SCA** model (right). The red dot in each saliency map indicates the location of the maximum value. Compared to traditional deep model, our method highlights both the semantic-aware top-down saliency and the contrast-aware bottom-up saliency.

Driven by the great success of deep neural networks, the deep learning scheme were recently introduced to saliency research and undoubtfully achieved ground-break success in predicting human eye-fixations (Kümmerer, Theis, and Bethge 2014; Huang et al. 2015; Pan et al. 2016). The remarkable performance of deep models relies very much on their ability of characterizing semantic information in the inputs. However, they do not explain the psychophysical evidence, and often produce unreasonable responses to obvious salient patterns. For instance, in Figure 1), the saliency maps produced by (Pan et al. 2016) only characterize the presence of meaningful objects, e.g. faces and fish, but fail to localize the most salient pattern in the image.

As discussed in (Bruce, Catton, and Janjic 2016), some eye fixations are directed at objects, while others are attracted by local feature contrast that is relatively detached from semantics. On one hand, the integration of deep neural network is inevitable, since traditional methods can not discover or make effective use of large-scale features that represent attractive semantic objects or interests. On the other hand, we must carefully formulate the model to ensure it is truely a model of saliency instead of a simplified object detector derived from recognition models.

Based on the above considerations, we proposed an integrated saliency model which combines both semantic-aware and contrast-aware saliency (namely, **SCA**) for effective eye fixation prediction. **SCA** contains two pathways and responses to both bottom-up and top-down cues. Compared to traditional deep model (Figure 1), our method highlights both the semantic interests and the bottom-up contrast. The main contributions of this paper are listed as follows:

- We propose an integrated saliency model (**SCA**) that characterizes both top-down and bottom-up saliency cues in a unified framework.
- The **SCA** model outperforms the state-of-art methods, including a recent deep model (winner of LSUN Chanllenge 2015), on all 5 eye-fixation dataset.
- Despite the superior performance in fixation prediction, the **SCA** model also produces plausible response to image with pre-attentive patterns and high-contrast objects.

Related Works

There are many excellent research works contributed to this topic, most of them can be found in recent surveys(Borji and Itti 2013; Li et al. 2014). In this section, we only introduce the related models used in our experiments, along with some recent deep models.

Traditional Saliency Model

(Itti, Koch, and Niebur 1998) implemented the very first computational model (**ITTI**) that generates 2D saliency map based on the Feature Integration theory (Treisman and Gelade 1980). The center-surround operation inspires many subsequent models (Gao, Mahadevan, and Vasconcelos 2008). Graph-based Visual Saliency (**GBVS**) model (Harel, Koch, and Perona 2006) adopts the same bio-inspired features in (Itti, Koch, and Niebur 1998) yet a different parallelized graph computation measurement to compute visual saliency.

Based on sparse representation, (Bruce and Tsotsos 2006) proposed the **AIM** model (Attention by Information Maximization) which adopts the self-information of ICA coefficients as the measure for signal saliency. Also based on information theory and sparse coding, (Hou and Zhang 2008) proposed the dynamic visual attention (**DVA**) model which defines spatio-temporal saliency as incremental coding length. (Garcia-Diaz et al. 2012) proposed the **AWS** (Adaptive Whitening Saliency) model which relies on a contextually adapted representation produced through adaptive whitening of color and scale features.

(Hou and Zhang 2007) proposed a highly efficient saliency detection algorithm by exploring the spectral residual (**SR**) in the frequency domain. **SR** highlights the salient regions by manipulating the amplitude spectrum of the images' Fourier transformation. As a theoretical revision of **SR**, (Hou, Harel, and Koch 2012) proposed a new visual descriptor named Image Signature (**SIG**) based on the Discrete Fourier Transform of images, which was proved to be more effective in explaining saccadic eye movements and change blindness of human vision.

Deep Saliency Model

Provided with enough training data, deep model can achieve ground-breaking performance that are far better than traditional methods, sometime even outperform humans. The ensembles of Deep Networks (**eDN**) (Vig, Dorr, and Cox 2014) is the first attemp at modeling saiency with deep models, which combines three different convnet layers using a linear classifer. Different from **eDN**, recent models such as **DeepGaze** (Kümmerer, Theis, and Bethge 2014), **SALICON** (Huang et al. 2015) and **FOCUS** (Bruce, Catton, and Janjic 2016) integrate pre-train layers from large-scale CNN models. Especially, (Huang et al. 2015) (**SALICON**), (Kruthiventi, Ayush, and Babu 2015) (**DeepFix**) and (Pan et al. 2016) (**DPN**) use the VGG net pre-trained on ImageNet to initialize their convolutional layers. Benefit from the powerfull visual representation embeded in VGG net, the above models significantly outperform traditional methods in eye-fxiation prediction task on almost all benchmark datasets. In this paper, we mainly compared our model with **DPN** (Pan et al. 2016) mainly because it is open source, easy to reimplemented, and has good efficiency as well as state-of-the-art performance.

The ground-breaking improvements achieved by integrating the pre-trained convolutional layers directly motivates our SCA model. However, we found that the fine-tuning of the deep network is very hard, and resulting model can not reliably detect and pop out high contrast signal in the input context, even with a large-scale training set such as SALICON (Jiang et al. 2015). In contrast, traditional adaptive methods such as AWS go through a pure bottom-up pathway and are hardly affected by the presence of semantics. This further inspires us to formulate our SCA model which handles top-down semantic cues and the bottom-up contrast in two independent pathways.

The Proposed Model

In this section, we first formulate our problem and then introduce our integrated saliency model **SCA**. We also present the implementation details of our model.

Problem Formulation

Let I be an image, and F be a set of pixel locations recording human eye fixations. The task of saliency prediction is to generate a saliency map which predicts the probability of every pixel in the image for being a true human eye fixation.

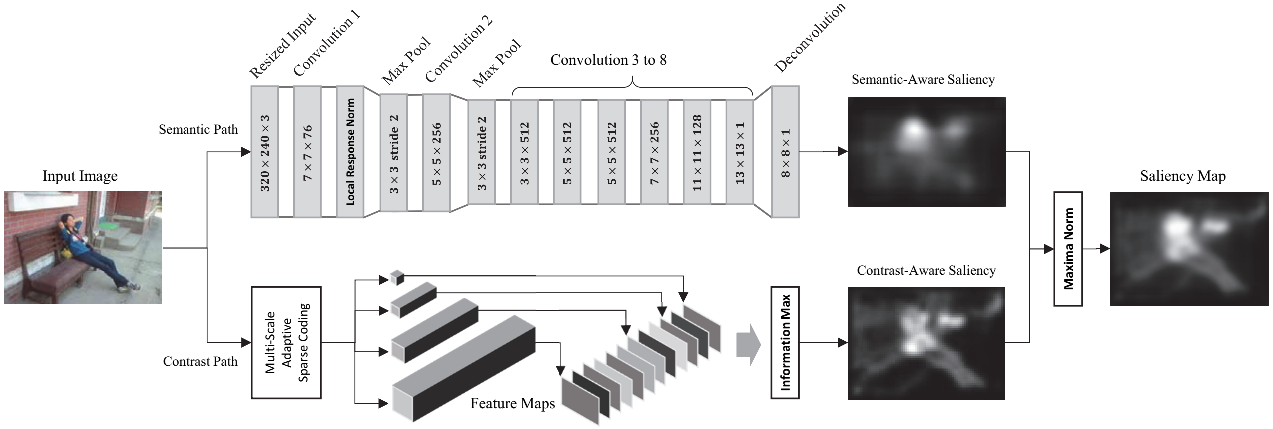

Figure 2: The Proposed **SCA** Saliency Model. **Top**: a deep neural network customized for semantic-aware saliency. **Bottom**: A workflow that discovers the high contrast salient patterns within the image context based on information maximization. The temporal outputs of the two pathways are integrated using maxima normalization.

Framework Overview

As shown in Figure 2, the proposed **SCA** model contains two pathways. The first pathway is a deep neural network customized for semantic-aware saliency, which aims to capture the semantic information in images, especially for the presence of meaningful objects and object parts. The second pathway is based on on-line feature learning and information maximization, which learns an adaptive representation for the input and discovers the high contrast patterns within the image context. The two pathways characterize both long-term and short-term attention cues and are integrated using maxima normalization.

Semantic-Aware Saliency

The first component of our integrated saliency model is a deep convnet customized for the computation of semantic-aware saliency (**SAS**). The **SAS** module is derived from **VGG** (Chatfield et al. 2014), an existing deep convnet originally trained for large-scale image classification. The **VGG** features learned from large-scale dataset such as ImageNet are highly correlated with the semantic information such as objects (deep-layer) or object parts (shallow-layer). There are many application that are based on the VGG feature, covering both low-level tasks like edge prediction and high-level problems such as video event detection.

Figure 2 shows the architecture of the **SAS** network. We follow a similar parameter setting of (Pan et al. 2016), which used the first 3 weight layers of **VGG** net, followed by a pooling layer and 6 convolutional layers. A deconvolution layer to obtain the semantic-aware saliency map that matches the size of the input. For any input image, we resize its size to $[240 \times 320]$ because humans can recognize most of the important objects in images at this resolution. Besides, a larger resolution means much more parameters and will significantly increase the training and testing time. The first 3 convolutional layers are initialized based on the pre-trained **VGG** net and the remaining are initialized randomly.

We used the data from **SALICON** dataset to train our **SAS** net. **SALICON** is currently the largest dataset available for saliency prediction task, which provides 10K images for training and 5K images for testing. Compared to traditional eye-fixation datasets, **SALICON** is much larger in size, which is more suitable for the training of deep convnet with large numbers of parameters. However, the fixation labels of **SALICON** is obtained from mouse-clicks instead of using eye-tracking devices, which might produce some small side-effects. We train our network based on 9K images from the training set, and use the rest 1K for validation. Standard preprocessing procedures are adopted including mean-subtraction and [-1,1] normalization on both the input images and the ground-truth saliency maps. For training, we adopted stochastic gradient descent with Euclidean loss using a batch size of 2 images for 24K iterations. L^2 weight regularizer was used for weight decay and the learning rate was halved after every 100 iterations.

The **SAS** pathway aims to capture the semantic information in images, especially for the presence of meaningful objects and object parts. The network is very light compared to those designed for object recognition (Chatfield et al. 2014) but much more complex than traditional feature-based saliency models.

Contrast-Aware Saliency

When too much semantic information is presented in the stimuli, e.g. lots of cars running by, human visual attention is more likely to be attracted by the stimulus with higher feature contrast, e.g. a car with the different color. In such scenarios, the influence of high-level semantics becomes less significant, and it's more reliable to use data-driven approaches for the estimation of saliency. This intuition leads to the second component of our **SCA** saliency model, namely contrast-aware saliency (**CAS**), which is based on multi-scale adaptive sparse representation and information maximization.

Multi-Scale Adaptive Sparse Feature Extraction Given an image I, we first turn it into a multi-scale patch-based representation $\{\mathbf{X}_i | i = 1, 2, 3, 4\}$ by scanning I with multiple sliding windows (window size $B_i \in \{1, 3, 5, 7\}$) from top-left to bottom-right. $\mathbf{X}_i$ is stored as a $M_i \times N_i$ matrix, where each column vector corresponds to a reshaped RGB image patch ($M_i = B_i \times B_i \times 3$ is the size of each patch, N_i is the total number of patches). For each $\mathbf{X}_i$, we apply Independent Component Analysis (**ICA**) (Hyvärinen and Oja 1997)[1] to generate a complete sparse dictionary $\mathbf{W}_i = [\mathbf{w}_1^i, \mathbf{w}_2^i, ..., \mathbf{w}_{M_i}^i]$, where each basis is an independent component of $\mathbf{X}_i$. Given $\mathbf{w}_j^i \in \mathbf{W}_i$ as the j-th basis at scale i , we can generate a feature vector $\mathbf{F}_j^i$ by treating $\mathbf{w}_j^i$ as a linear filter:

$$\mathbf{F}_j^i = {\mathbf{w}_j^i}^{\mathbf{T}} \mathbf{X}_i \tag{1}$$

where $\mathbf{F}_j^i(k)$ denotes the response value of k-th patch for the j-th basis at scale i.

Saliency by Information Maximization Inspired by (Bruce and Tsotsos 2006), we measure the contrast-aware saliency of each patch by the self-information of its multi-scale sparse feature. Specifically, the CAS value of the k-th patch at scale i is defined as:

$$\begin{aligned} \mathbf{S}_i(k) &= -\log \prod_j p_{i,j}(\mathbf{F}_j^i(k)) \\ &= -\sum_j \log p_{i,j}(\mathbf{F}_j^i(k)), \end{aligned} \tag{2}$$

where $p_{i,j}(.)$ is the probability density function of the j-th feature at scale i, which can be estimated using histogram method. We can obtain multiple saliency maps focusing on different salient patterns at various scales. The final **CAS** map is obtained by summing up all single-scale **CAS** maps.

Figure 3 shows some visual examples of **SAS** and **CAS** maps on natural images, which clearly illustrates their preference and highlight in the given image context. The **SAS** highlights semantics such as faces and body parts, while the **CAS** map emphasizes those visual contents with high feature contrast.

Saliency Integration

Fusing saliency detection results of multiple models has been recognized as a challenging task since the candidate models are usually developed based on different cues or assumptions. Fortunately, in our case, the integration problem is relatively easier since we only consider the outputs from two pathways. Since there is no prior knowledge or other top-down guidance can be used, it's safer to utilize the map statistics to determine the importance of each pathway. Intuitively in the final integration stage, we combine the results from two pathways by summing them after Maxima Normalization (**MN**) (Algorithm 1).

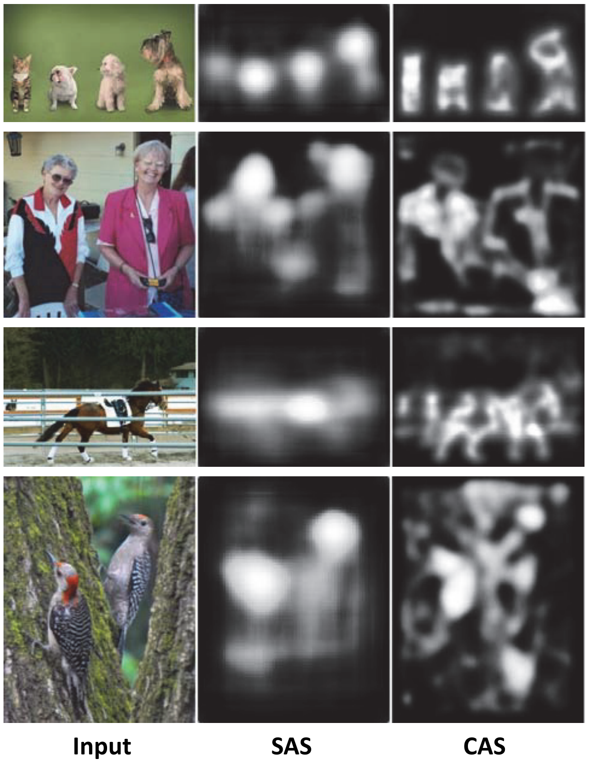

Figure 3: Examples of **SAS** and **CAS** maps. **SAS** highlights semantics while **CAS** emphasizes the contextual contrast.

Algorithm 1 Maxima Normalization $N_{max}(S, t)$

```
Input: 2D intensity map $S$, thresh of local maxima $t = 0.1$
Output: Normalized Saliency Map $S_N$
Set the number of maxima $N_M = 0$
Set the sum of the maxima $V_M = 0$
Set Global Maxima $G_M = \max(S)$
for all pixel $(x, y)$ of $S$ do
    if $S(x, y) > t$ then
        $R = \{S(i, j) | i = x - 1, x + 1, j = y - 1, y + 1\}$
        if $S(x, y) > \max(R)$ then
            $V_M = V_M + S(x, y)$
            $N_M = N_M + 1$
        end if
    end if
end for
$S_N = S \cdot (G_M - V_M / N_M)^2 / G_M$
return Normalized map $S_N$
```

The Maxima Normalization operator $N_{max}(.)$ was originally proposed for the integration of conspicuous maps from multiple feature channels (Itti, Koch, and Niebur 1998), which has been demonstrated very effective and has a very convincing psychological explanation. In addition to the MN integration strategy, we also test two alternative methods in the experiment : Average-Pooling (**AP**) and Max-Pooling (**MP**)(Wang et al. 2016).

[1] **FastICA**: https://research.ics.aalto.fi/ica/fastica/

Experiments

We evaluate our model in two tasks: 1) prediction of human eye fixation and 2) response to artificial images with controlled salient patterns.

Datasets and Evaluation Metric

For the eye-fixation prediction task, we used 6 public available dataset including: Bruce (Bruce and Tsotsos 2006), Cerf (Cerf et al. 2008) , imgSal (Li et al. 2013), Judd (Judd et al. 2009), PASCAL-S (Li et al. 2014) and SALICON (Jiang et al. 2015). Each dataset contains a group of natural images and the corresponding eye-fixations captured from human subjects. Note that, among the 6 datasets, SALICON is the largest in size but its eye-fixation data is labeled by mouse clicks instead of using eye-tracking devices. Thus in our experiment, we use SALICON to train our semantic-aware saliency module and apply the rest 5 dataset for performance evaluation. Table 1 show some basic statistics of the 6 dataset.

DataSet	Images	Subjects	Device
Bruce	120	20	Eye-Tracker
Cerf	200	7	Eye-Tracker
ImgSal	235	21	Eye-Tracker
Judd	1003	15	Eye-Tracker
PASCAL	850	8	Eye-Tracker
SALICON	15.000	16	Mouse-Click

Table 1: Statistics of the 6 eye-fixation dataset.

Many evaluation metrics have been developed to measure the quality of saliency maps in predicting human eye-fixations. Following the procedure of the recent benchmarking surveys (Borji and Itti 2013; Li et al. 2014), we use the shuffled Area Under ROC curve (**sAUC**) (Tatler, Baddeley, and Gilchrist 2005) as our major evaluation metric in our experiment. The original **AUC** metric scores a saliency map according to its ability to separate the eye fixations from random points. In **sAUC**, positive samples are taken from the eye-fixation of the test image, whereas the negative ones are sampled from other images. We use the GPU implementation of **sAUC** from (Li et al. 2014) for the experiments. Post-processing parameters (e.g. blur kernel) are also optimized for each method to ensure a fair comparison.

In addition to the eye-fixation datasets, we also test the models on a small group of artificial images consisting of controlled salient patterns. Most of the images are made based on solid evidence, such as preattentive feature, found in psychological experiments. Thus, we can further evaluate the plausibility of a model by checking whether it can produce a reasonable response to such images.

Results on Eye-Fixation Prediction

We have compared our integrated model with several classic saliency prediction algorithms including: **ITTI** (Itti, Koch, and Niebur 1998), **AIM** (Bruce and Tsotsos 2006), **GBVS** (Harel, Koch, and Perona 2006), **DVA** (Hou and Zhang 2008), **SUN** (Zhang et al. 2008), **SIG** (Hou, Harel, and Koch 2012), and **AWS** (Garcia-Diaz et al. 2012). We've also compared to a recent CNN model, **DPN** (Pan et al. 2016), which is a very efficient and effective model based on deep convolutional network. Since our model is partially trained on SALICON, we mainly report the evaluation results on the other 5 datasets. We show some visual examples of the tested saliency models in Figure4.

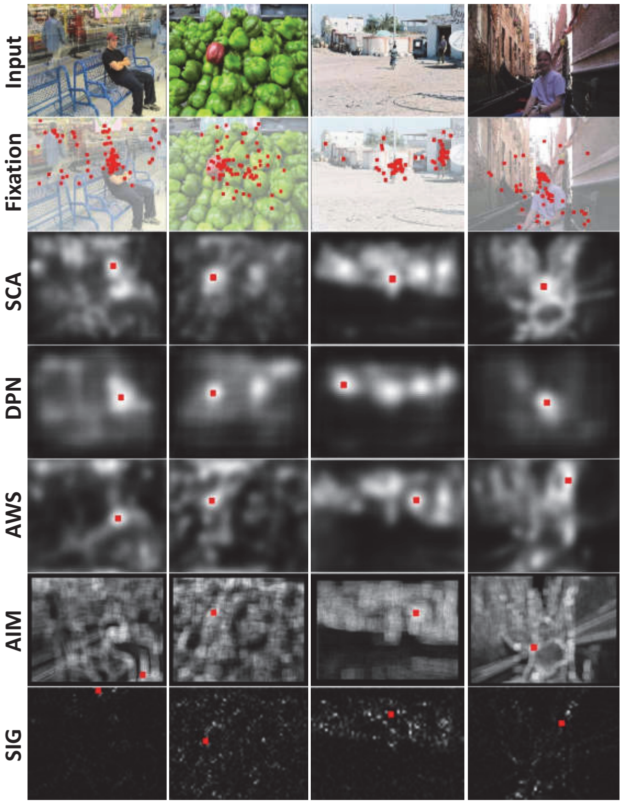

Figure 4: Visual comparisons of different saliency models. The red dot in each saliency map indicates the location of the maximum saliency. The proposed **SCA** can effectively locate the most salient pattern in the input.

Figure 5 shows the main **sAUC** results. We show the performance curves (bottom) of the test models over different blur kernel σ and use bar chart (top) to show the **sAUC** score of each model under its optimal σ. Note that , the proposed **SCA** method achieves the best **sAUC** performance on all 5 datasets. It's also important to note that the deep model **DPN** and our **CCA** model significantly outperform traditional approaches on two largest datasets, **Judd** and **PASCAL**, because both models apply deep network to embed semantics into the saliency computation process.

Table 2 shows the optimal **sAUC** score of the three alternative integration strategy, including **AP** (average pooling), **MP** (max pooling) and the default **MN** (maxima normalization). Overall, **SCA-MN** performs the best, but only slightly better than **SCA-AP**. **SCA-MN** and **SCA-AP** are much better than **SCA-MP**.

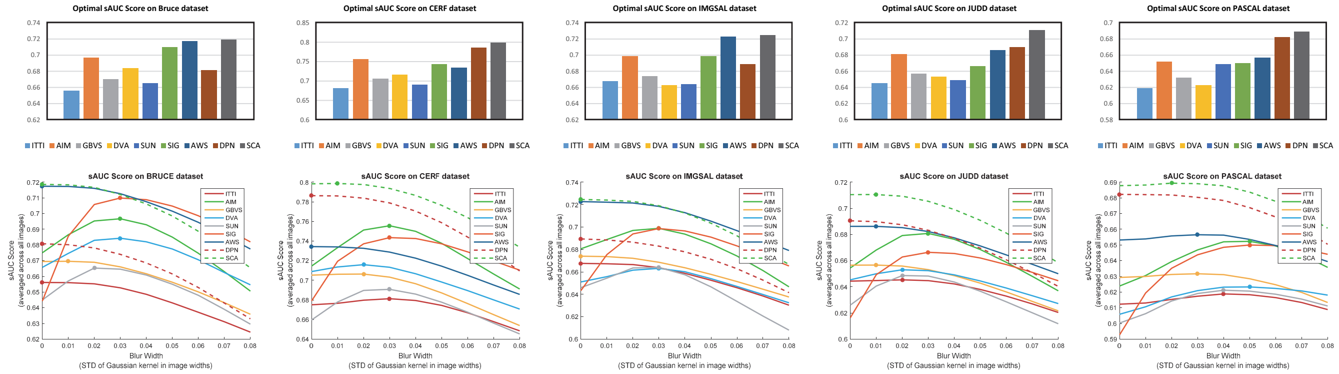

Figure 5: Performance for eye-fixation prediction on 5 datasets. **Top**: Bars showing **sAUC** score under the optimal σ of each model. **Bottom**: Curves depicting **sAUC** score as a function of a blurring kernel σ.

	ITTI	**AIM**	**GBVS**	**DVA**	**SUN**	**SIG**	**AWS**	**DPN**	**SCA-AP**	**SCA-MP**	**SCA-MN**
Bruce	0.656	0.697	0.670	0.684	0.665	0.710	0.717	0.681	**0.719**	0.718	**0.719**
Cerf	0.681	0.756	0.706	0.716	0.691	0.743	0.734	0.786	0.798	0.787	**0.799**
ImgSal	0.668	0.699	0.674	0.663	0.664	0.699	0.723	0.689	**0.726**	0.725	0.725
Judd	0.645	0.681	0.657	0.653	0.649	0.666	0.686	0.690	0.710	0.705	**0.711**
PASCAL	0.619	0.652	0.632	0.623	0.649	0.650	0.657	0.682	0.688	0.685	**0.689**
Weighted **Avg.**	0.6416	0.6795	0.6546	0.6502	0.6547	0.6722	0.6849	0.6946	0.7116	0.7074	**0.7123**

Table 2: The optimal **sAUC** score on all eye-fixation datasets. We show **SCA** with three integration strategy: Average Pooling (**SCA-AP**), Max Pooling (**SCA-MP**) and Maxima Normalization (**SCA-MN**)

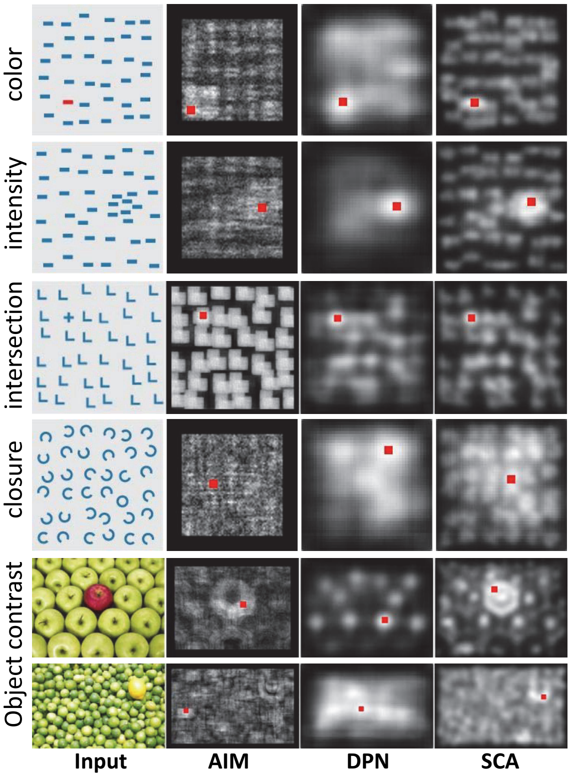

Figure 6: Response to artifical salient patterns. The red dot in each map indicates the location of the maximum saliency.

Response to Artifical Salient Patterns It has been well accepted that the response to the artificial patterns adopted in attention related experiments can indicate the biological plausibility of the tested models. In Figure 6, we show some example results based on the images provided by (Hou and Zhang 2007) and (Bruce, Catton, and Janjic 2016). The red dot in each map indicates the location of the maximum saliency. In comparison to the results of **AIM** and **DPN**, our **SCA** model successfully locates all salient patterns in these images, which further demonstrate its effectiveness and psychophysical plausibility.

Conclusion

In this paper, we proposed an integrated model of both semantic-aware and contrast-aware saliency (**SCA**) combining both bottom-up and top-down cues for effective eye fixation prediction. As the middle results, the **SAS** and **CAS** maps generated by the two pathways clearly show their preference and highlight in the given image. Experimental results on 5 benchmark dataset and artificial images demonstrate the superior performance and better plausibility of the proposed **SCA** model over both classic approaches and the recent deep model.

Acknowledgement

This research was partially supported by the Discovery Project of Australian Research Council, DP130103252, FT120100718 and FT130101530.

References

Alexe, B.; Deselaers, T.; and Ferrari, V. 2012. Measuring the objectness of image windows. *IEEE TPAMI* 34(11):2189–2202.

Borji, A., and Itti, L. 2013. State-of-the-art in visual attention modeling. *IEEE TPAMI* 35(1):185–207.

Bruce, N., and Tsotsos, J. 2006. Saliency based on information maximization. In *NIPS'06*, 155–162.

Bruce, N. D.; Catton, C.; and Janjic, S. 2016. A deeper look at saliency: Feature contrast, semantics, and beyond. In *IEEE CVPR'16*, 516–524.

Cerf, M.; Harel, J.; Einhäuser, W.; and Koch, C. 2008. Predicting human gaze using low-level saliency combined with face detection. In *NIPS'08*, 241–248.

Chatfield, K.; Simonyan, K.; Vedaldi, A.; and Zisserman, A. 2014. Return of the devil in the details: Delving deep into convolutional nets. *arXiv preprint arXiv:1405.3531*.

Frintrop, S. 2010. General object tracking with a component-based target descriptor. In *Robotics and Automation (ICRA), 2010 IEEE International Conference on*, 4531–4536. IEEE.

Gao, D.; Mahadevan, V.; and Vasconcelos, N. 2008. On the plausibility of the discriminant center-surround hypothesis for visual saliency. *Journal of Vision* 8(7):13–13.

Garcia-Diaz, A.; Leboran, V.; Fdez-Vidal, X. R.; and Pardo, X. M. 2012. On the relationship between optical variability, visual saliency, and eye fixations: A computational approach. *Journal of Vision* 12(6):17–17.

Harel, J.; Koch, C.; and Perona, P. 2006. Graph-based visual saliency. In *NIPS'06*, 545–552.

Hou, X., and Zhang, L. 2007. Saliency detection: A spectral residual approach. In *IEEE CVPR'07*, 1–8.

Hou, X., and Zhang, L. 2008. Dynamic visual attention: searching for coding length increments. In *NIPS'08*, 681–688.

Hou, X.; Harel, J.; and Koch, C. 2012. Image signature: Highlighting sparse salient regions. *IEEE TPAMI* 34(1):194–201.

Huang, X.; Shen, C.; Boix, X.; and Zhao, Q. 2015. Salicon: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *IEEE ICCV'15*, 262–270.

Hyvärinen, A., and Oja, E. 1997. A fast fixed-point algorithm for independent component analysis. *Neural Compuptation* 9:1483–1492.

Itti, L.; Koch, C.; and Niebur, E. 1998. A model of saliency-based visual attention for rapid scene analysis. *IEEE TPAMI* 20(11):1254–1259.

Itti, L. 2004. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE TIP* 13(10):1304–1318.

Jiang, M.; Huang, S.; Duan, J.; and Zhao, Q. 2015. Salicon: Saliency in context. In *IEEE CVPR'15*, 1072–1080. IEEE.

Judd, T.; Ehinger, K.; Durand, F.; and Torralba, A. 2009. Learning to predict where humans look. In *IEEE ICCV'09*, 2106–2113. IEEE.

Koch, C., and Ullman, S. 1987. Shifts in selective visual attention: towards the underlying neural circuitry. In *Matters of Intelligence*. Springer. 115–141.

Kruthiventi, S. S.; Ayush, K.; and Babu, R. V. 2015. Deepfix: A fully convolutional neural network for predicting human eye fixations. *arXiv preprint arXiv:1510.02927*.

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

Leboran, V.; Garcia-Diaz, A.; Fdez-Vidal, X.; and Pardo, X. 2016. Dynamic whitening saliency. *IEEE TPAMI*.

Li, J.; Levine, M. D.; An, X.; Xu, X.; and He, H. 2013. Visual saliency based on scale-space analysis in the frequency domain. *IEEE TPAMI* 35(4):996–1010.

Li, Y.; Hou, X.; Koch, C.; Rehg, J. M.; and Yuille, A. L. 2014. The secrets of salient object segmentation. In *IEEE CVPR'14*, 280–287.

Pan, J.; McGuinness, K.; Sayrol, E.; O'Connor, N.; and Giro-i Nieto, X. 2016. Shallow and deep convolutional networks for saliency prediction. In *IEEE CVPR'16*, 598–606.

Rosenholtz, R.; Dorai, A.; and Freeman, R. 2011. Do predictions of visual perception aid design? *ACM Transactions on Applied Perception (TAP)* 8(2):12.

Salah, A. A.; Alpaydin, E.; and Akarun, L. 2002. A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition. *IEEE TPAMI* 24(3):420–425.

Siagian, C., and Itti, L. 2009. Biologically inspired mobile robot vision localization. *Robotics, IEEE Transactions on* 25(4):861–873.

Sun, X.; Yao, H.; Ji, R.; and Liu, X.-M. 2014. Toward statistical modeling of saccadic eye-movement and visual saliency. *IEEE TIP* 23(11):4649–4662.

Tatler, B.; Baddeley, R.; and Gilchrist, I. 2005. Visual correlates of fixation selection: Effects of scale and time. *Vision Research* 45(5):643–659.

Treisman, A. M., and Gelade, G. 1980. A feature-integration theory of attention. *Cognitive Psychology* 12(1):97–136.

Vig, E.; Dorr, M.; and Cox, D. 2014. Large-scale optimization of hierarchical features for saliency prediction in natural images. In *IEEE CVPR'14*, 2798–2805.

Wang, J.; Borji, A.; Kuo, C.-C. J.; and Itti, L. 2016. Learning a combined model of visual saliency for fixation prediction. *IEEE TIP* 25(4):1566–1579.

Zhang, L.; Tong, M. H.; Marks, T. K.; Shan, H.; and Cottrell, G. W. 2008. Sun: A bayesian framework for saliency using natural statistics. *Journal of Vision* 8(7):32–32.