This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2905607, IEEE Transactions on Pattern Analysis and Machine Intelligence

Inferring Salient Objects from Human Fixations

Wenguan Wang, *Member, IEEE*, Jianbing Shen, *Senior Member, IEEE*,
Xingping Dong, Ali Borji, and Ruigang Yang, *Senior Member, IEEE*

Abstract—Previous research in visual saliency has been focused on two major types of models namely fixation prediction and salient object detection. The relationship between the two, however, has been less explored. In this work, we propose to employ the former model type to identify salient objects. We build a novel Attentive Saliency Network (ASNet) that learns to detect salient objects from fixations. The fixation map, derived at the upper network layers, mimics human visual attention mechanisms and captures a high-level understanding of the scene from a global view. Salient object detection is then viewed as fine-grained object-level saliency segmentation and is progressively optimized with the guidance of the fixation map in a top-down manner. ASNet is based on a hierarchy of convLSTMs that offers an efficient recurrent mechanism to sequentially refine the saliency features over multiple steps. Several loss functions, derived from existing saliency evaluation metrics, are incorporated to further boost the performance. Extensive experiments on several challenging datasets show that our ASNet outperforms existing methods and is capable of generating accurate segmentation maps with the help of the computed fixation prior. Our work offers a deeper insight into the mechanisms of attention and narrows the gap between salient object detection and fixation prediction.

Index Terms—Image saliency, salient object detection, fixation prediction, deep learning.

1 Introduction

Salient object detection (SOD) has been studied extensively for over a decade since the work in [2]. It has been shown effective in a wide range of applications such as object segmentation [3], [4], visual tracking [5], object proposal generation [6], person re-identification [7], and image resizing [8], [9], just to mention some representative ones.

Recently, the use of deep neural networks for saliency detection has been trending. Although promising results have been achieved, connections between these models and how humans explicitly choose salient objects or watch natural scenes are less clear. This largely limits the interpretation capability of the SOD models. Moreover, as discussed in [10], [11], the majority of current SOD datasets are biased as they often contain one or two salient objects in images, compared to human eye-tracking datasets with several objects in natural scenes. Given a complex cluttered scene containing several objects (such as the ones shown in Figure 1 (a)), current SOD models occasionally fail to detect the most salient object.

In this paper, we take a further step towards a more biologically inspired SOD model. We are mainly motivated by the behavioral studies (*e.g.* [12], [13]) that have investigated how humans select salient objects explicitly and how these judgments relate to eye movements during scene free viewing. Unlike some other works (*e.g.* [14], [15], [16]), here we do not aim to mimic low-level mechanisms of attention in the brain or use them in our model. Our model benefits from a high-level prior of *fixation map*. The human attention prior, represented by eye movements or from a fixation prediction model, is instinctive and more consistent with visual processing of human visual system in natural scene free-viewing. The suggested model not only generates high-quality object saliency maps, but also pushes the boundary of SOD research by building a close connection to human fixation prediction (FP). As shown in Figure 1, our model infers object saliency (Figure 1 (b)) using the fixation prior (Figure 1 (c)), where this prior acts as a selective mechanism to enhance the saliency representation for the purpose of accurate object saliency inference. Such a fixation prior imitates human visual attention mechanisms and allows the suggested model to explicitly segment out the most visually important object(s) in an interpretable manner.

(a) (b) (c)

Fig. 1. Given complex scenes like (a), what are the salient objects? We propose the Attentive Saliency Network (ASNet) that infers the object saliency (b) from predicted fixation maps (c), which is consistent with human visual attention mechanisms.

Our algorithm is based on the core views of previous studies [12], [13] which explored the relationship between eye movements (implicit saliency) and explicit object

- *W. Wang is with Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology, and also with Inception Institute of Artificial Intelligence, UAE. (Email: wenguanwang.ai@gmail.com)*
- *J. Shen and X. Dong are with Beijing Laboratory of Intelligent Information Technology, School of Computer Science, Beijing Institute of Technology (Email: shenjianbing@bit.edu.cn).*
- *A. Borji is with the Center for Research in Computer Vision, University of Central Florida.*
- *R. Yang is with the University of Kentucky, Lexington, KY 40507.*

- *A preliminary version of this work has appeared in CVPR 2018 [1].*

- *Corresponding author: Jianbing Shen*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2905607, IEEE Transactions on Pattern Analysis and Machine Intelligence

saliency. These studies confirmed a strong correlation between fixations and salient objects. Further, in contrast to traditional saliency models that treat FP and SOD as two separate tasks, the suggested model integrates these two tasks in an interconnected and comprehensive way. The fixation map provides a high-level signal, which is learned from upper layers of our neural network. It is then used for SOD in a top-down manner. This process is straightforward and similar to how humans process a scene sequentially (*i.e.*, first paying attention to important areas of a scene quickly, and then taking more efforts for precise segmentation). Additionally, leveraging the rich information from existing large-scale eye movement datasets can improve the robustness and generalization ability of SOD models.

The proposed Attentive Saliency Network (ASNet) is based on a variant of recurrent neural network: *convolutional LSTM* (convLSTM) [17], which has convolutional structures in both the input-to-state and state-to-state transitions. As opposed to the *fully connected LSTM*, convLSTM encodes the spatial information via convolution operations. This is crucial for pixel-wise saliency labeling. Another key advantage of LSTM lies on iterative removal of irrelevant information and learning powerful representations through updating the memory cell. This allows the ASNet to progressively optimize features for better saliency estimation in a feed-forward manner. By stacking multiple convLSTMs, the ASNet is trainable to gradually compute object saliency from fixation map in a top-down manner [18].

Several evaluation metrics have been employed for assessing the quality of SOD model predictions with respect to groundtruth labels. These metrics, however, often do not agree with each other making fair model comparison very challenging. This motivates us to develop a set of new loss functions for the SOD task. These loss functions, derived from current widely-adopted evaluation metrics, encourage the ASNet to generate better results through accounting for different characteristics of evaluation factors of SOD. The proposed ASNet is fully convolutional and has no need of any pre-processing procedure. It has a fast processing speed of 15 fps (on one GPU; w/o IO time and CRF).

1.1 Contributions

Our contribution in this paper is four-fold:

1) **A unified and biologically inspired deep SOD model that emphasizes learning object saliency from visual attention prior.** The suggested model effectively infers salient objects (captured in lower network layers) from the fixation map (encoded in higher layers), which is more consistent with human visual attention behavior. This goes one step beyond previous deep learning based saliency models and offers a deep insight into the confluence between fixation prediction and salient object detection.
2) **A novel network architecture that applies a stack of convLSTMs for iterative saliency feature learning and refinement.** We present the Attentive Saliency Network (ASNet) which is a hierarchy of convLSTMs for inferring object saliency in a step-by-step, coarse-to-fine, and top-down manner. ConvLSTM has the advantage of the improved flow of information with recurrent connections, which results in more powerful saliency representation.
3) **A set of essential loss functions for SOD.** We propose new SOD loss functions, derived from existing SOD evaluation metrics, for capturing several quality factors. As we will show empirically, these new loss functions lead to higher performance.
4) **Empirical evidence is provided, offering a deeper insight into the relation between SOD and FP tasks.** Experiments on SOD benchmarks confirm that object-level saliency can benefit from fixation analysis. This helps relate saliency for fixation prediction and object detection and encourages future work in this direction.

These contributions altogether bring both effectiveness and efficiency to our proposed SOD model. As described in Section 4, the thorough evaluations on several popular benchmarks clearly show the advantage of our algorithm in comparison with state-of-the-art solutions.

This paper builds upon our recent conference paper [1] and extends it in several ways. First, we provide a more in-depth discussion of the proposed algorithm, including motivations, network structures and implementation. Second, we offer an exhaustive and insightful overview of the recent work on SOD and FP, and present a more insightful discussion regarding the relationship between these two tasks. Third, more ablation studies are designed for a thorough examination of our model. Fourth, we report extensive experimental results with an additional SOD dataset. Last but not least, based on our experiments, we draw several important conclusions, which are expected to inspire future follow-up works in this direction.

1.2 Organization

The rest of the paper is organized as follows. In Section 2, we first provide a short overview of the most notable works for SOD and FP tasks, introduce previous studies regarding the relations between SOD and FP, and present detailed discussions for typical network architecture designs for SOD and FP. Then, in Section 3, we illuminate the proposed ASNet in details, which is capable of inferring object-level saliency from visual attention prior. In Section 4, we offer both quantitative and qualitative experimental analysis of the proposed algorithm over various public benchmarks. The massive experimental results clearly demonstrate that our ASNet compares favorably with the state-of-the-art. Finally, concluding remarks can be found in Section 5.

2 Related Work

In this section, we first briefly review the fixation prediction (Section 2.1) and salient object detection literature (Section 2.2). Then, in Section 2.3, we introduce studies exploring the relationship between the two tasks and discuss the representative network architectures of current deep saliency models.

2.1 Fixation Prediction (FP)

Fixation Prediction (FP) aims to predict where humans fixate during scene free viewing. It has a long history from [19] and is still an active area in vision research.

Early attention models [20], [21], [22] were mainly based on stimulus-driven bottom-up mechanisms and cognitive assumptions about visual attention. They typically leverage biologically-inspired features (contrasts in intensity, color, or orientation) and allocate high fixation probabilities towards the regions which are noticeably different from their neighborhoods over these visual features. From a computational standpoint, these attention models can be further classified into several categories [23], such as cognitive [19], [24], Bayesian [25], decision theoretic [26], information theoretic [27], graphical [20], spectral analysis [28], pattern classification [29], *etc*. We refer the readers to [23], [30] for more detailed reviews.

More recently, many **deep learning based visual attention models** have been proposed. As one of the earliest such models, the eDN model [31] presented an architecture that automatically learns deep representations for predicting fixations. Authors searched optimal deep features from a family of hierarchical neuromorphic networks (each individual is a small and shallow CNN with a maximum of 3 layers) and fed the features into an SVM for saliency prediction. Following this work, a wide variety of deep learning schemes has emerged [32], [33], [34], [35], [36], [37], [38]. Different from eDN that learns deep features from scratch, subsequent works fine-tuned existing deeper neural networks (*e.g*., VGG-16 [39]) on eye-tracking dataset and thus gained improved performance [32], [35]. Huang *et al*. [33] built a multi-stream saliency model, known as SALICON, for emphasizing multi-scale saliency representation learning. Similarly, Liu *et al*. [34] constructed an ensemble of CNNs, termed Multiresolution-CNN (Mr-CNN), where each of these CNNs is trained to classify image patches, at a particular scale, for saliency. Another recent work, DVA [36], proposed to fuse multi-layer saliency responses within one single network, instead of learning multiple network streams with different input scales. In [40], Jetley *et al*. tested several loss functions based on probability distance measures and found that the Bhattacharyya distance gives the best performance. With the availability of large-scale eye-tracking data and the strong learning ability of deep neural network, those deep learning solutions achieved significantly better performance, compared with traditional non-deep saliency techniques.

2.2 Salient Object Detection (SOD)

Salient object detection (SOD) aims at highlighting salient object regions in images. Different from FP that is originated from cognitive and psychology research communities, SOD is a computer vision task driven by object-level applications [42], [43]. The history of SOD task is comparatively short and can be traced back to the works of Liu *et al*. [2] and Achanta *et al*. [44]. Most of **non-deep learning SOD models** [45], [46], [47] are based on low-level features and rely on certain heuristics (*e.g*., *color contrast* [3], *background prior* [48]). For obtaining uniformly highlighted salient objects and clear object boundaries, an over-segmentation process that generates regions [49], super-pixels [50], [51], [52], [53], or object proposals [54] is often integrated into above models. Please see [13] for a comprehensive overview.

Most recently, **deep learning based SOD models** have made substantial improvement. A few early methods were based on fully-connected networks and leveraged image segmentations (over-segmentation [55], superpixels [56], [57], [58], or object proposals [59]) as basic processing units. They fed deep features extracted from the image segments into a fully-connected classifier for saliency score prediction. Although improved performance was achieved over previous non-deep learning SOD models, these fully-connected SOD schemes fall short in learning essential spatial information well and are quite time-consuming as they need to process all image segments one by one. To overcome this limitation, latest SOD models [41], [60], [61], [62], [63], [64] are largely built upon fully convolutional neural networks (FCN), which well emphasizes end-to-end spatial saliency representation learning and leads to faster saliency prediction within only one feed-forward process. For example, Wang *et al*. [65] proposed a stage-wise scheme that progressively optimizes saliency predictions with the assistance of a heuristic saliency prior. Hou *et al*. [66] introduced a skip-layer structure that concatenates features from all different network layers together for a more comprehensive fusion of high-level semantic and low-level detailed saliency representations. In [60], an SOD model called Amulet, was proposed to aggregate multi-level feature maps at each resolution predict saliency in a recursive manner. Islam *et al*. [67] built an encoder-decoder network that emphasizes multiple salient object detection, relative ranking, and subitizing, simultaneously. Some other models utilized multi-level representations [61], incorporated level set into saliency learning [62], or exploited more complex network architectures [68], [69], [70], [71], [72].

2.3 Relationship between FP and SOD

Although SOD has been extensively studied in computer vision research, only few studies [10], [12], [13] have explored how humans explicitly choose salient objects. They quantitatively confirmed that object saliency judgments agree with human eye movements. According to the analyses in [11], [13], there exists a strong correlation between explicit saliency judgments and free-viewing fixations, which can be viewed as two proxies of visual attention. Borji *et al*. [10] further pointed out that an SOD model should involve two steps: 1) selecting objects to process, and 2) segmenting the object area. Unfortunately, current SOD models made a decent effort towards the second step, while largely ignoring the first challenge. Li *et al*. [12] studied the connections between the two tasks via a series of experimental analyses over popular FP and SOD datasets. They demonstrated that, unlike FP datasets, there exists a heavy bias in many widely used SOD datasets. Most of the SOD datasets have only few obvious objects in the scene. Motivated by above studies, we build a visual attention driven SOD model, where the fixation map offers an explicit explanation for the choice of salient objects. In this way, it investigates the properties of salient objects from the perspective of human visual attention mechanisms and brings an in-depth exploration of the relationship between where people look in scenes and what they choose as salient objects.

Next, we discuss several recent representative deep models in SOD or FP from the view of network architecture. This allows us to better situate our work with respect to previous studies and helps highlight our contributions. Here,

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2905607, IEEE Transactions on Pattern Analysis and Machine Intelligence

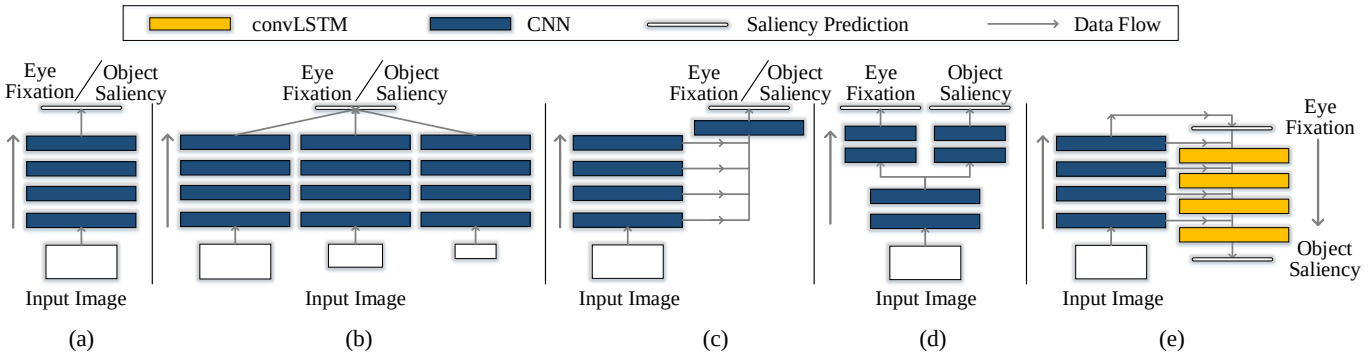


Fig. 2. (a)-(c) Typical network architectures used in previous FP or SOD models (from left to right): single-stream network, multi-stream network, and skip-layer network. (d) Branched network adopted in [41], where FP and SOD are achieved via two branches sharing several bottom layers. (e) The adopted ASNet captures fixation map from upper layers, which is indicative of the inference of object saliency from lower layers. Stack of convLSTMs are adopted for iteratively optimizing features, while preserving spatial information. See Section 2.3 for a more detailed discussion.

we mainly consider fully convolutional neural network based saliency models, as they have been the mainstream in recent years.

As illustrated in Figure 2, most deep learning models for FP or SOD only consider one single task. Typical architectures can be formalized into three categories, namely *single-stream network*, *multi-stream network*, and *skip-layer network*. More specifically, *single-stream network* [32], [40], [35], [70], [62] is a standard convolutional neural network architecture composed of a sequential cascade of repeated convolutional layers, intermediated by pooling and activation operations. A typical single-stream learning architecture is illustrated in Figure 2 (a). *Multi-stream network* [33], [56], [55], [34], as depicted in Figure 2 (b), typically has multiple network streams, where each individual stream is trained with input at a particular resolution. It explicitly learns multi-scale saliency features. The outputs from different network streams are combined together for producing the final saliency prediction. *Skip-layer network* [66], [57], [60], [36], [61] concatenates multi-layer responses for producing the final output. The core scheme of skip-layer network is shown in Figure 2 (c). This kind of network design is to integrate hierarchical features inside a neural network for capturing distinctive high-level objectness and detailed low-level information simultaneously. All in all, although different types of network architectures have been employed, previous deep learning based works often treat FP and SOD as two unrelated tasks.

Different from previous saliency models that perform FP and SOD separately, we exploit the correlation between fixations and salient objects via tightly coupling these two tasks in a unified deep learning architecture. So far, only few methods consider FP and SOD tasks jointly. In [64], fixation map from a *pre-trained* FP model is leveraged as an extra feature for SOD. They did not emphasize learning both FP and SOD simultaneously. In [41], FP and SOD are achieved via two separate network branches, which only share weights in several lower layers (Figure 2 (d)). In our algorithm, as illustrated in Figure 2 (e), fixation map (high-level knowledge captured in top layers) is used for guiding accurate object-level saliency estimation in lower layers. A stack of convLSTMs (the yellow rectangles in Figure 2 (e)) is further built for coarse-to-fine, step-wise saliency refinement with gradual utilization of more detailed spatial information from lower layers. Thus, our approach goes beyond above work by learning FP and SOD within a unified network and in a top-down end-to-end manner.

3 Our Approach

Given an input image, the goal is to produce a pixel-wise saliency map to highlight salient object regions. As demonstrated in Figure 3, the proposed ASNet first captures a global and high-level understanding of a scene in its higher layers, by learning to predict human fixations (Section 3.1). Afterwards, it uses a stack of convLSTMs to progressively infer object saliency from the fixation map in a top-down and coarse-to-fine manner (Section 3.2). The whole network is simultaneously trained to predict fixation locations and to detect salient objects in an end-to-end manner (Section 3.3).

3.1 Fixation Predicting

At the bottom of ASNet resides a stack of convolutional layers where the lower layers respond to primitive image features such as edges, corners and shared common patterns, and the higher layers extract semantic information such as objects or faces. The ASNet learns the FP as a high-level task towards modeling human visual attention mechanism with the utilization of features from higher layers, and achieves the SOD by optimizing the fixation prior with the features from the lower layers.

The lower convolutional layers are borrowed from the first five convolutional blocks of VGGNet [39] (13 convolutional layers in total). Each of the convolutional blocks is followed by a max-pooling layer with downsampling stride 2. We omit the last pooling layer (*pool5*) for preserving more spatial information. For a training image, with a typical resolution of $224\times224\times3$, we compute a convolutional layer (with 1 channel) by applying a 3×3 kernel with the *sigmoid* activation function, to the last convolutional feature map ($14\times14\times512$). The result is a probability map $P \in [0,1]^{14\times14}$ which is used as a fixation prior from global and high-level image context. The model for the task of FP is trained via minimizing the following *Kullback-Leibler Divergence* (KL-Div) loss function:

$$\mathcal{L}_{Att}(G,P) = \frac{1}{14\times14}\sum_{x}^{14\times14} g_x \log(\frac{g_x}{p_x}), \qquad (1)$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2905607, IEEE Transactions on Pattern Analysis and Machine Intelligence

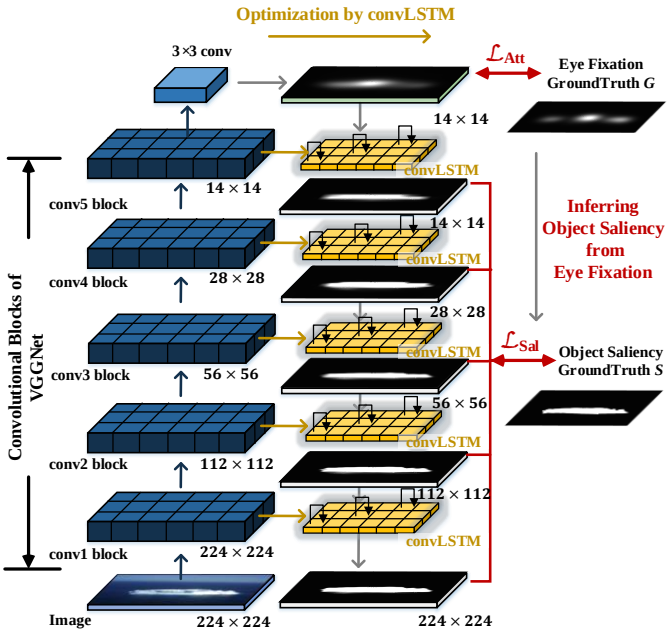


Fig. 3. Architecture of the proposed ASNet. The fixation map is learned from the upper layers and is used by the ASNet to locate the salient objects. Then, the fine-grained object-level saliency is gradually inferred from lower layers and is successively optimized via the recurrent architecture of convLSTM. Zoom-in for details.

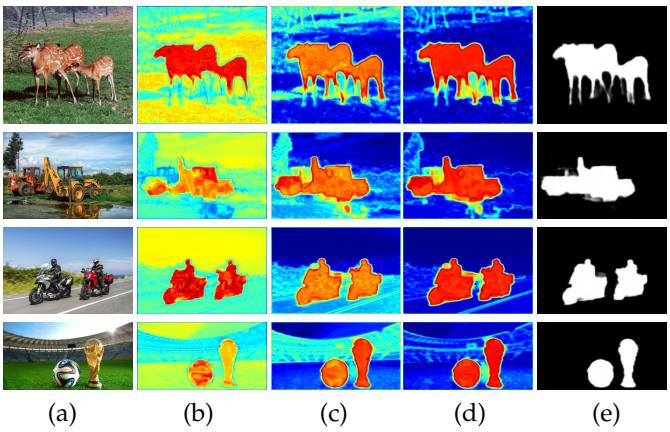


Fig. 4. Illustration of utilizing the recurrent mechanism of convLSTM for iteratively optimizing saliency representation. The figures in columns (b)-(d) correspond to the learned saliency features $\mathcal{H}_1, \mathcal{H}_2, \mathcal{H}_3$ in different time steps: $t = \{1, 2, 3\}$, and (e) shows the final salient object prediction maps. We observe that the convLSTM is able to gradually improve the learned saliency representations via suppressing the responses from the background and enhancing the foreground features step by step. Please see Section 3.2 for more details.

where G denotes the resized ground-truth attention map $G \in [0, 1]^{14\times14}$ and $g_x \in G$, $p_x \in P$. The gray-scale fixation map is obtained via filtering the binary fixation map using a Gaussian filter with small variance. The KL-Div measure, the minimization of which is equivalent to cross-entropy minimization, is widely used in visual saliency prediction [33], [36]. In the next section, we will leverage such fixation map as the prior for producing object-level saliency.

3.2 Detecting Object-Level Saliency with Fixation Prior

The fixation map P gives a coarse but informative prior regarding visually salient regions. A number of previous studies for pixel-labeling tasks such as semantic segmentation [73], and SOD [69], [65], have shown that neural networks are capable of producing fine-gained labeling results via incorporating high-level information encoded in upper network layers with finer features represented at lower layers. Here, we desire our model to be able to infer precise object-level saliency from the fixation map predicted in the upper network layers in a top-down, coarse-to-fine manner.

The network is trained for detecting and successively refining the salient objects via aggregating information from high-level fixation map and the spatially rich information from low-level network features. To be specific, as shown in Figure 3, the SOD is computed in a top-down fashion, successively integrating information from earlier layers. Multiple convLSTM networks [17] (the yellow blocks in Figure 3) are stacked for building more meaningful feature representations with recurrent connections. We leverage the sequential nature of LSTM to process features in an iterative way (see Figure 4). For a certain layer, convLSTM discards less informative features while enhances useful features, thus generating gradually improved saliency maps. In this way, each convLSTM takes the preceding saliency estimation with finer scale representations as input and carries out a sequence of iterative optimization operations to generate a refined salient object prediction map.

ConvLSTM extends traditional fully connected LSTM [74] to consume spatial features. Basically, this is achieved by substituting dot products with convolutional operations in the LSTM equations. ConvLSTM has convolutional structures in both the input-to-state and state-to-state transitions, which can preserve the spatial information of convolutional feature map, thus enabling our network to produce a pixel-wise labeling.

A schematic diagram of convLSTM is presented in Figure 5. Similar to traditional gated LSTMs, the convLSTM uses the memory cells and gates to control information flow. It works by sequentially updating an internal state $\mathcal{H}$ and memory cell $\mathcal{C}$, according to the values of three sigmoid gates i, f, c. At each step t, as a new input $\mathcal{X}_t$ arrives, its information will be accumulated to the cell if the input gate i_t is activated. Also, the past cell status $\mathcal{C}_{t-1}$ could be "forgotten" in this process if the forget gate f_t is on. Whether the latest cell status $\mathcal{C}_t$ should be propagated to the final state $\mathcal{H}_t$ is further controlled by the output gate o_t. Formally, above memory update process at step t is driven by the following equations:

$$i_t = \sigma(W_i^{\mathcal{X}} * \mathcal{X}_t + W_i^{\mathcal{H}} * \mathcal{H}_{t-1} + b_i), \quad (2)$$

$$f_t = \sigma(W_f^{\mathcal{X}} * \mathcal{X}_t + W_f^{\mathcal{H}} * \mathcal{H}_{t-1} + b_f), \quad (3)$$

$$o_t = \sigma(W_o^{\mathcal{X}} * \mathcal{X}_t + W_o^{\mathcal{H}} * \mathcal{H}_{t-1} + b_o), \quad (4)$$

$$\mathcal{C}_t = f_t \circ \mathcal{C}_{t-1} + i_t \circ \tanh(W_c^{\mathcal{X}} * \mathcal{X}_t + W_c^{\mathcal{H}} * \mathcal{H}_{t-1} + b_c), \quad (5)$$

$$\mathcal{H}_t = o_t \circ \tanh(\mathcal{C}_t), \quad (6)$$

where '$*$' denotes the convolution operator and '$\circ$' represents element-wise product. σ and $\tanh$ are the activation functions of logistic sigmoid and hyperbolic tangent. The inputs $\mathcal{X}_t$, cell memory $\mathcal{C}_t$, hidden states $\mathcal{H}_t$ and gates

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2905607, IEEE Transactions on Pattern Analysis and Machine Intelligence

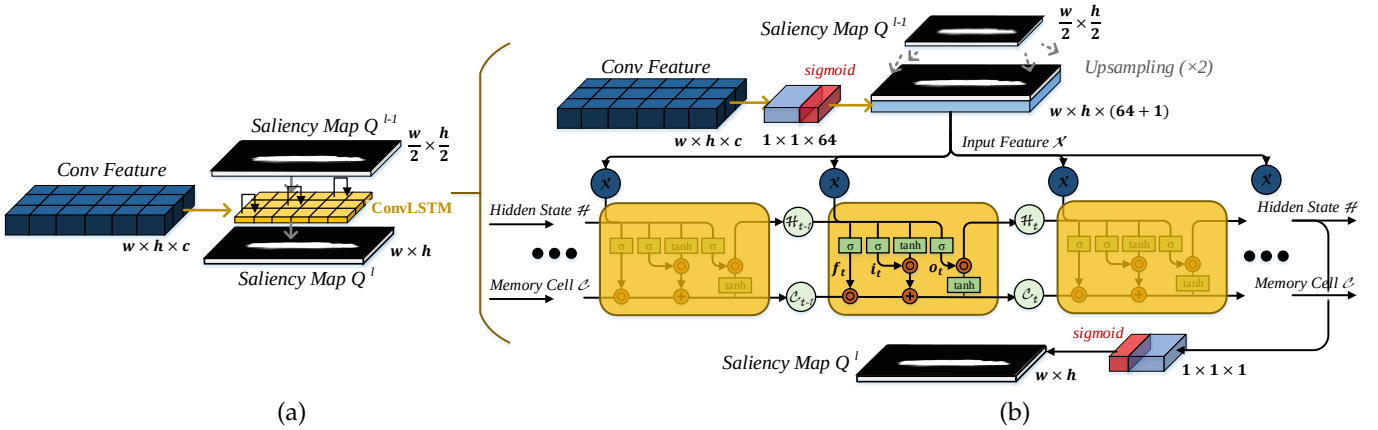


Fig. 5. Illustration of our convLSTM based object-level saliency optimization, where (b) shows detailed architecture of our convLSTM optimization module in (a). Zoom-in for details.

i_t, f_t, c_t are 3D tensors whose spatial dimensions are the same. Ws and bs are the learned weights and biases.

In our case, the convLSTM takes the features $\mathcal{X}$ extracted from the convolutional neural network (from the last convolutional layers prior to pooling layers) as input, and produces refined saliency features for final saliency estimation. Since it operates on static images, the input features in all steps are the same: $\mathcal{X}_1 = \ldots = \mathcal{X}_t = \mathcal{X}$ (see Figure 5). Here, we take the advantage of the recurrent nature of LSTM for iteratively optimizing the saliency features of static images, instead of using LSTM for modeling the temporal dependency of sequential data. Figure 4 presents an intuitive illustration of our rationale of utilizing convLSTM, where (b)-(d) show the saliency features $\mathcal{H}$s learned in different steps are gradually improved. The noisy responses from the background are successfully suppressed and the saliency features are iteratively and gradually enhanced, thanks to the utilization of the recurrent mechanism of convLSTM.

With the learned fixation prior $P \in [0,1]^{14\times14}$, we combine P with the convolutional features from *conv5-3* and feed them into a convLSTM. In each time step, the convLSTM is trained to infer the salient object with the knowledge of fixation information, and to sequentially optimize the features with the updated memory cell and hidden states (see Figure 5 (b)). Thus, the features are reorganized towards better representation of the object-level saliency. More specifically, we first compress the feature responses from *conv5-3* layer via a convolutional layer with 64 filters to lower computational costs and adopt the *sigmoid* activation for regularizing the response from features to lie within the same range ($[0,1]$) of P. Then, the attention prior map P is concatenated with the compressed features along the channel dimension and fed into the convLSTM. We apply a 1×1 convolution kernel to the final convLSTM output $\mathcal{H}$ for obtaining an object-level saliency map $Q \in [0,1]^{14\times14}$.

Several different metrics have been proposed for evaluating saliency models and no single metric can fully summarize the performance of a model. This motivates us to combine the classical *weighted cross-entropy* loss function with *precision*, *recall*, *F-measure*, and *MAE* metrics for more efficient training. Given the ground-truth salient object annotation S (here $S \in \{0,1\}^{14\times14}$ for *conv5-3* layer), the overall loss function is defined as:

$$\begin{aligned}\mathcal{L}_{\text{Sal}}(S,Q) = \mathcal{L}_{\text{C}}(S,Q) + \alpha_1\mathcal{L}_{\text{P}}(S,Q) + \alpha_2\mathcal{L}_{\text{R}}(S,Q) \\ + \alpha_3\mathcal{L}_{\text{F}}(S,Q) + \alpha_4\mathcal{L}_{\text{MAE}}(S,Q),\end{aligned} \tag{7}$$

where αs are balancing parameters and are empirically set as $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = 0.1$. $\mathcal{L}_{\text{C}}$ is the *weighted cross-entropy* loss function, which is widely adopted for training SOD models and used as the primary loss in our case:

$$\begin{aligned}\mathcal{L}_{\text{C}}(S,Q) = \frac{1}{N}\sum\nolimits_x \big(\gamma \cdot (1-s_x)\cdot \log(1-q_x) \\ + (1-\gamma)\cdot s_x \cdot \log q_x\big),\end{aligned} \tag{8}$$

where N is the total number of pixels and $s_x \in S$, $q_x \in Q$. γ refers to the ratio of salient pixels in ground truth S. *Weighted cross-entropy* loss handles the imbalance between number of salient and non-salient pixels.

$\mathcal{L}_{\text{P}}$, $\mathcal{L}_{\text{R}}$ and $\mathcal{L}_{\text{F}}$ are computed similar to *precision*, *recall* and *F-measure* scores:

$$\mathcal{L}_{\text{P}}(S,Q) = -\sum\nolimits_x s_x \cdot q_x \Big/ \big(\sum\nolimits_x q_x + \epsilon\big), \tag{9}$$

$$\mathcal{L}_{\text{R}}(S,Q) = -\sum\nolimits_x s_x \cdot q_x \Big/ \big(\sum\nolimits_x s_x + \epsilon\big), \tag{10}$$

$$\mathcal{L}_{\text{F}}(S,Q) = -\frac{(1+\beta^2)\cdot\mathcal{L}_{\text{P}}(S,Q)\cdot\mathcal{L}_{\text{R}}(S,Q)}{\beta^2\cdot\mathcal{L}_{\text{P}}(S,Q) + \mathcal{L}_{\text{R}}(S,Q) + \epsilon}. \tag{11}$$

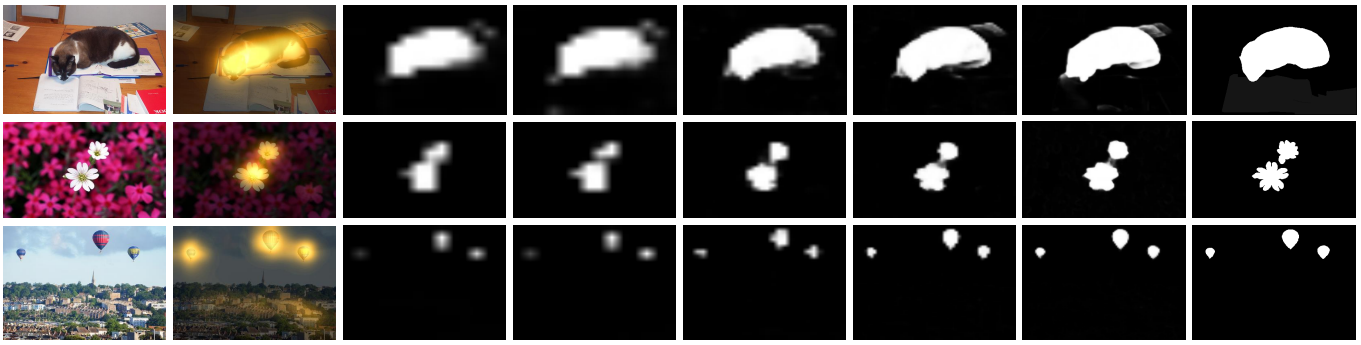
where $\beta^2 = 0.3$ as suggested by [44], and ϵ is a regularization constant. Since *precision*, *recall* and *F-measure* are similarity metrics and higher values are better, negative values are used for minimization.

$\mathcal{L}_{\text{MAE}}$ is derived from the *mean absolute error* (*MAE*) measure that computes the discrepancy between the saliency map Q and the ground-truth map S:

$$\mathcal{L}_{\text{MAE}}(S,Q) = \frac{1}{N}\sum\nolimits_x |s_x - q_x|. \tag{12}$$

After obtaining the object-level saliency map $Q \in [0,1]^{14\times14}$ inferred from the fixation map P, we upsample ($\times2$) Q and feed it to the next convLSTM with the compressed features ($28\times28\times64$) from *conv4-3* layer for more detailed refinement. Above process is iteratively applied to *conv4-3*, *conv3-3*, *conv2-2* and *conv1-2* layers, respectively. This brings a top-down coarse-to-fine learning scheme. Finally, the ASNet outputs a high-quality object-level saliency

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2905607, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) Image | (b) Visual Attention Prior | (c) SOD Result from *conv-5* | (d) SOD Result from *conv-4* | (e) SOD Result from *conv-3* | (f) SOD Result from *conv-2* | (g) SOD Result from *conv-1* | (h) Groundtruth for SOD

Fig. 6. Illustration of visual attention driven, coarse-to-fine object saliency inference process. It can be observed that the visual attention prior (b) predicted by the top layer of ASNet is able to guide fine-grained object-level saliency estimation in lower layers in a top-down fashion. With incorporating finer features from different conv blocks of VGG-16 base network, the SOD results (c)-(g) can be gradually optimized in a coarse-to-fine manner (best viewed in color). Please see Section 3.2 for more details.

mask (224×224×1). Figure 6 provides an illustration of above top-down, coarse-to-fine saliency estimation process. It can be observed that the proposed ASNet is capable to gradually improve the saliency estimation via integrating finer features from the lower network layers. In sum, the ASNet is able to effectively infer the object saliency thanks to 1) a learnable fixation prior, 2) progressively updating saliency features with a recurrent architecture and 3) efficiently merging spatially rich information from lower layers in a top-down manner.

3.3 Implementation Details

Overall loss: Let $\mathcal{I}=\{I_k \in \mathbb{R}^{224\times224\times3}, k=1,\ldots,K\}$ denote all the training images (resized into 224×224) borrowed from existing SOD or FP datasets. Since there are only few datasets that offer annotations for both SOD and FP tasks, most of the training images are either labeled with human fixation annotations or object-level saliency masks. Let $y_k^A \in \{0,1\}$ and $y_k^S \in \{0,1\}$ indicate whether we have the attention annotation G_k and object saliency mask S_k for the k-th training image. Our final loss function can be expressed as:

$$\mathcal{L}=\sum_{k=1}^{K} y_k^A \cdot \mathcal{L}_{\text{Att}}(G_k,P_k)+\sum_{k=1}^{K} y_k^S \cdot \sum_{\ell=1}^{5} \mathcal{L}_{\text{Sal}}(S_k^\ell, Q_k^\ell), \tag{13}$$

where the loss functions $\mathcal{L}_{\text{Att}}$ and $\mathcal{L}_{\text{Sal}}$ are defined in Equation 1 and Equation 7, respectively. The indicators y_k^A and y_k^S are employed to remedy missing ground truth in corresponding tasks. That is the error is not propagated back when the annotations are not offered. The $\ell \in \{1,\ldots,5\}$ refers to the ℓ-th convLSTM with *conv*-ℓ block in ASNet. With the hierarchical loss functions, each layer in ASNet has direct access to the gradients from the loss function leading to an implicit deep supervision [75]. We set the time steps to 3 in our convLSTM and employ 3×3 kernels in convolution operations.

Training datasets: Another advantage of ASNet is that it can use data from both SOD and FP benchmarks. We consider three large-scale saliency datasets: SALICON [76], THUS10K [3], and DUT-OMRON [50]. The **SALICON** dataset is widely used in the domain of FP. The **THUS10K** dataset, containing 10,000 natural images, is commonly used for SOD. These two datasets have annotations for fixations and salient objects, respectively. We further utilize the **DUT-OMRON** dataset which has 5,168 challenging images with annotations for both FP and SOD. The fixation maps were generated from the eye-tracking data of 5 subjects during a 2-second viewing. Detailed descriptions of these datasets can be found in Table 1. To study the correlation between human fixation and explicit saliency judgment on DUT-OMRON dataset, we follow the protocol in [11] to use the segmentation mask to explain the fixation map. More specifically, during the computation of AUC-Judd metric (see the detailed definition in Section 4.1.2), human fixations are considered as the positive set and some points sampled from other non-fixation positions as the negative set. The segmentation mask is then used as a binary classifier to separate positive samples from negative samples. The correlation score 0.767 (*std* = 0.098) is significantly above chance (0.5) using t-test ($p < 0.05$), showing the strong correlation between human dynamic visual attention and salient object determination. A visual illustration of the consistency between SOD and FP can be found in Figure 7; showing again that visual attention significantly tends to fall in the salient object area.

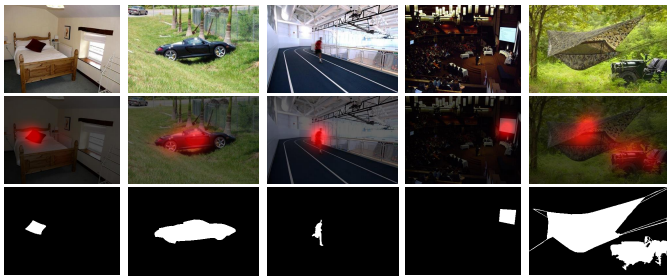


Fig. 7. Illustration of the consistency between FP and SOD annotations of the training datasets. From top to bottom: image, FP annotation and SOD groundtruth.

Training settings: All training images are uniformly resized into 224×224 pixels. In each training iteration, we use a min-batch of 10 images, which is sampled from above 3

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2905607, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 1
Statistics of the datasets used for training and testing the ASNet.

Aspect	Dataset	Publication	Year	#Images	Annotation		#Viewers	Resolution
					FP	SOD		
Training	[1] SALICON [76]	CVPR	2015	15,000	✓		-	640×480
	[2] THUS10K [3]	CVPR	2011	10,000		✓	-	$\max(w,h) = 400$
	[3] DUT-OMRON [50]	CVPR	2013	5,168	✓	✓	5	$\max(w,h) = 400$
Testing	[4] PASCAL-S [12]	CVPR	2014	850	✓	✓	8	$\max(w,h) = 500$
	[5] MIT1003 [29]	ICCV	2009	1,004	✓		15	$\max(w,h) = 1024$
	[6] ECCSD [49]	CVPR	2013	1,000		✓	-	$\max(w,h) = 400$
	[7] HKU-IS [55]	CVPR	2015	4,447		✓	-	$\max(w,h) = 400$
	[8] SOD [77]	CVPR workshop	2010	300		✓	-	321×481

[1] http://salicon.net/ [2] https://mmcheng.net/zh/msra10k/ [3] http://saliencydetection.net/dut-omron/
[4] http://cbi.gatech.edu/salobj/ [5] http://people.csail.mit.edu/tjudd/WherePeopleLook/
[6] http://www.cse.cuhk.edu.hk/leojia/projects/hsaliency/dataset.html
[7] https://i.cs.hku.hk/~gbli/deep_saliency.html [8] http://elderlab.yorku.ca/SOD/

datasets and ensure data balance between SOD and FP. For THUS10K [3] dataset, data augmentation techniques (*e.g.*, flipping, rotation) are also adopted. Our model is implemented in Python in Keras, and trained with the Adam optimizer [78]. Pre-trained VGGNet is used to initialize the convolutional layers in the base network (*i.e.*, the *conv1* to *conv5* block). The parameters of other layers are randomly assigned. During training, the learning rate is set to 0.0001 and is decreased by a factor of 10 every two epochs. The networks were trained for 12 epochs with early stopping strategy. The entire training procedure takes about 10 hours with a NVIDIA TITAN X GPU and a 4.0 GHz Intel processor.

Testing phase: During testing, we feed the testing image $I \in \mathbb{R}^{224\times224\times3}$ to the ASNet. With a feed-forward process, we obtain both eye fixation prediction map $P \in [0,1]^{14\times14}$ and a set of salient objects prediction maps $\{Q^\ell\}_{\ell=1}^5$, where Q^ℓ obtained from the ℓ-th convLSTM with *conv*-ℓ block. The most fine-grained object-level saliency prediction $Q^1 \in [0,1]^{224\times224}$ is used as our final SOD result. Since our ASNet does not need other prior information such as heuristic saliency prior or pre-processing steps such as superpixel over-segmentation, it achieves a fast processing speed of 15 fps (without using CRF), which is faster than most deep learning based contenders.

4 Experimental Results

In this section, we present both quantitative and qualitative experiments for thoroughly assessing the performance of the proposed ASNet. Specifically, in Section 4.1, we provide details of our experimental settings, including testing datasets, evaluation metrics, *etc*. Then, in Section 4.2.1, we first examine the performance of ASNet for the FP task, using PASCAL-S [12] and MIT1003 [29] datasets. The goal of this experiment is to investigate the effectiveness of the learned fixation map prior, instead of comparing it with the state-of-the-art FP models. Furthermore, in Section 4.2.2, we evaluate the performance of the ASNet for the primary SOD task, using four public benchmarks, namely PASCAL-S [12], ECCSD [49], HKU-IS [55] and SOD [77], compared to 19 state-of-the-art SOD models. The experiments clearly demonstrate the robustness, effectiveness, and efficiency of our algorithm. To better understand the contributions of different aspects of our proposed algorithm, in Section 4.3, we implement several variants of our method to conduct detailed ablative studies.

4.1 Experimental Setup

4.1.1 Testing datasets

Five datasets including PASCAL-S [12], MIT1003 [29], ECCSD [49], HKU-IS [55] and SOD [77] are utilized for testing our model. All of these datasets are available online. Table 1 presents statistics of these datasets.

PASCAL-S includes 850 natural images with multiple complex objects derived from the validation set of the PASCAL VOC 2012 [79]. For each image, fixations during 2 seconds of 8 subjects are offered, and the salient object annotation is generated according to the fixation data. Since PASCAL-S offers both annotations for FP and SOD, we report the evaluation results for both tasks over this dataset.

MIT1003 is a representative benchmark for FP. It contains totally 1,003 images from Flicker and LabelMe, including 779 landscape and 228 portrait images. The ground-truth saliency maps were generated from eye-tracking data of 15 human observers.

ECCSD is a typical SOD dataset, which contains 1000 natural images with pixel-accurate ground truth annotations. These images generally contain semantically meaningful but structurally complex salient objects.

HKU-IS is also widely used for SOD. It has 4447 images which are selected by meeting at least one of the following three criteria: multiple overlapping salient objects, objects touching the image boundary and low color contrast.

SOD contains 300 images borrowed from the Berkeley Segmentation Dataset (BSD) [80]. Seven subjects were employed for identifying the object(s) they perceived as the most salient.

4.1.2 Evaluation metrics

For the FP task, there are several ways to measure the agreement between model predictions and eye movements. Previous studies on saliency metrics [81] show that it is difficult to achieve a fair comparison of saliency models using any single metric. Here, we carried out our quantitative experiments by comprehensively considering a variety of different metrics, including Normalized Scanpath Saliency

TABLE 2
Details of the evaluation metrics used for the FP task. See Section 4.1.2 for more details.

Evaluation Metrics	Category	Groundtruth
CC	Distribution-based	Continuous Saliency Map G
SIM	Distribution-based	Continuous saliency Map G
NSS	Location-based	Discrete fixation Map F
AUC-Judd	Location-based	Discrete fixation Map F
shuffled AUC	Location-based	Discrete fixation Map F

(NSS), Similarity Metric (SIM), Linear Correlation Coefficient (CC), AUC-Judd, and shuffled AUC. These metrics are selected since they are widely-accepted and standard for evaluating saliency models. Above saliency evaluation metrics can be roughly categorized into two categories: location-based and distribution-based metrics [81]. The first category (*i.e.*, s-AUC, AUC-Judd, NSS) considers saliency maps at discrete fixation locations, while the second category (*i.e.*, CC, SIM) treats both ground-truth fixation maps and predicted saliency maps as continuous distributions. For the sake of simplification, in the following section, we denote the predicted saliency map as P, the map of fixation locations as F and the continuous saliency map (distribution) as G. In Table 2, we list the characteristics of our adopted evaluation metrics. Next, we describe these evaluation metrics in detail.

Normalized Scanpath Saliency (NSS) is a metric specifically designed for saliency map evaluation [82]. Given a saliency map P and a binary map of fixation locations F:

$$\begin{aligned} \text{NSS} &= \frac{1}{N}\sum\nolimits_x \overline{P}(x) \times F(x), \\ N &= \sum\nolimits_x F(x), \quad \overline{P} = \frac{P - \mu(P)}{\sigma(P)}, \end{aligned} \tag{14}$$

where N is the total number of eye positions and $\sigma(\cdot)$ stands for standard deviation. This metric is calculated by taking the mean of scores assigned by the unit normalized saliency map (with zero mean and unit standard deviation) at fixations.

Linear Correlation Coefficient (CC) is a statistical method generally used for measuring how correlated or dependent two variables are. CC can be used to interpret saliency and fixation maps, P and G, as random variables to measure the linear relationship between them:

$$\text{CC} = \frac{cov(P, G)}{\sigma(P) \times \sigma(G)}, \tag{15}$$

where $cov(P, G)$ is the covariance of P and G. It ranges between -1 and +1, and a score close to -1 or +1 indicates a perfect alignment between the two maps.

Similarity Metric (SIM) is also known as the histogram intersection metric. It measures the similarity between two distributions, viewed as histograms [83]. SIM is computed as the sum of the minimum values at each pixel location, after normalizing the input maps:

$$\begin{aligned} \text{SIM} &= \sum\nolimits_x \min(P'(x), G'(x)), \\ \sum\nolimits_x P'(x) &= 1, \quad \sum\nolimits_x G'(x) = 1, \end{aligned} \tag{16}$$

where P' and G' are normalized to be probability distributions, given a saliency map P and the continuous fixation map G. An SIM of one indicates that the distributions are the same, while an SIM of zero indicates no overlap.

AUC, defined as the area under the receiver operating characteristic (ROC) curve, is widely used to evaluate the maps estimated by saliency models [27]. Given an image and its groundtruth fixation points, fixated points and other points are regarded as the positive and negative sets, respectively. Then, the computed saliency map is binarized into salient and non-salient regions by using a threshold. Through varying the threshold from 0 to 1, the ROC curve is obtained by plotting true positive rate versus false positive rate, with its underneath area calculated as the AUC score. AUC can be greatly influenced by center-bias and border cut. Depending upon the choice of the non-fixated distribution, there are several variants of AUC. In our experiments, we adopt the AUC-Judd (AUC-J) [83], and the shuffled AUC (s-AUC) [25]. The former variant chooses non-fixated points with a uniform distribution, while the latter, shuffled AUC, uses human fixations of other images in the dataset as the non-fixated distribution.

For the SOD task, three standard metrics, namely PR-curve, F-measure, and MAE, are employed for quantitative evaluation.

For **PR-curve**, given a saliency map with continuous values normalized to the range of 0 and 255, we first compute the corresponding binary maps by using every possible fixed integer threshold. Let B denote the binary mask corresponding to a continuous saliency map S using a threshold, and G indicate the groundtruth mask. The precision and recall are computed as: precision $= |B \cap G|/|B|$, and recall $= |B \cap G|/|G|$, respectively, where $|\cdot|$ accumulates the non-zero entries in a mask. Then, we compute the precision/recall pairs of all binary maps to plot the PR curve by a mean value over all saliency maps in a given dataset.

F-measure is formulated by a weighted combination of precision and recall:

$$\text{F-measure} = \frac{(1 + \beta^2) \times \text{precision} \times \text{recall}}{\beta^2 \times \text{precision} + \text{recall}}, \tag{17}$$

where β^2 is set to be 0.3 to weigh precision more than recall as suggested in [44]. The reason for emphasizing precision more than recall is that recall rate is not as important as precision. For instance, as demonstrated in [2], [13], 100% recall can be easily achieved by setting the whole image as foreground.

The definition of **MAE score** can be found in Equation 13. MAE score offers a direct way for measuring the dissimilarity between saliency estimate and the groundtruth.

4.2 Performance of ASNet

4.2.1 Performance on FP task

Compared FP models: We evaluated the fixation prior map generated by the ASNet compared to 12 state-of-the-art fixation models, including 5 classical models: Itti's model (ITTI) [19], Graph-Based Visual Saliency (GBVS) [20], Attention based on Information Maximization (AIM) [27], Boolean Map based Saliency (BMS) [21], Context-Aware Saliency (CAS) [22], and 7 deep learning based models:

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2905607, IEEE Transactions on Pattern Analysis and Machine Intelligence

TABLE 3
Quantitative comparison of different FP models on the MIT1003 [29] dataset. The best scores are marked in **bold**. See Section 4.2.1 for more details.

Methods	AUC-J ↑	SIM ↑	s-AUC ↑	CC ↑	NSS ↑
Mr-CNN [34]	0.80	0.35	0.73	0.38	1.36
SALICON [33]	0.85	0.42	0.74	0.53	1.86
DVA [36]	0.87	0.50	**0.77**	0.64	2.38
†DVA [36]	0.87	0.51	**0.77**	**0.65**	**2.40**
Shallow-Net [35]	-	-	0.68	-	1.60
Deep-Net [35]	0.86	0.40	0.73	0.51	1.73
†Deep-Net [35]	0.87	0.42	0.74	0.50	1.80
SU [41]	-	-	0.73	-	2.08
eDN [31]	0.85	0.30	0.66	0.41	1.29
BMS [21]	0.79	0.33	0.69	0.36	1.25
CAS [22]	0.76	0.32	0.68	0.31	1.07
AIM [27]	0.79	0.27	0.68	0.26	0.82
GBVS [20]	0.83	0.36	0.66	0.42	1.38
ITTI [19]	0.77	0.32	0.66	0.33	1.10
ASNet-14×14	0.87	0.49	0.73	0.60	2.01
ASNet-28×28	**0.88**	**0.52**	0.75	**0.65**	2.30

- The authors in [41], [35] have not released detailed results.
† indicates the models retrained on SALICON [76] and DUT-OMRON [50] datasets.

TABLE 4
Quantitative comparison of different FP models on the PASCAL-S [12] dataset. The best scores are marked in **bold**. See Section 4.2.1 for more details.

Methods	AUC-J ↑	SIM ↑	s-AUC ↑	CC ↑	NSS ↑
Mr-CNN [34]	0.79	0.34	0.71	0.40	1.35
SALICON [33]	0.86	0.46	0.72	0.58	1.88
DVA [36]	0.89	0.52	0.76	0.66	2.26
†DVA [36]	0.89	0.53	**0.77**	0.67	2.30
Shallow-Net [35]	-	-	0.69	-	1.90
Deep-Net [35]	0.87	0.42	0.71	0.55	1.74
†Deep-Net [35]	0.88	0.44	0.72	0.58	1.90
SU [41]	-	-	0.73	-	2.22
eDN [31]	-	-	0.65	-	1.42
BMS [21]	0.79	0.34	0.67	0.39	1.28
CAS [22]	0.78	0.34	0.67	0.36	1.12
AIM [27]	0.77	0.30	0.65	0.32	0.97
GBVS [20]	0.84	0.36	0.65	0.45	1.36
ITTI [19]	0.82	0.36	0.64	0.42	1.30
ASNet-14×14	**0.90**	0.55	0.74	0.70	2.26
ASNet-28×28	**0.90**	**0.59**	0.74	**0.73**	**2.43**

- The authors in [35], [41], [31] have not released detailed results.
† indicates the models retrained on SALICON [76] and DUT-OMRON [50] datasets.

ensemble of Deep Networks (eDN) [31], Saliency in Context (SALICON) [33], Saliency Unified (SU) [41], Multi-resolution Convolutional Neural Networks (Mr-CNN) [34], Deep Visual Attention (DVA) [36], Shallow attention Network (Shallow-Net) [35] and Deep attention Network (Deep-Net) [35]. In addition, to pursue a more complete and fair comparison, we retrain DVA and Deep-Net on both SALICON [76] and DUT-OMRON [50] datasets. Results are reported over MIT1003 [29] and PASCAL-S [12] datasets.

Quantitative FP results: Our ASNet is able to generate a fixation prediction map P from top layer (using the feature from *conv5-3* layer; see Figure 3), which is relatively rough, and much smaller (only 14×14) compared to existing fixation models. For the sake of a fairer comparison and deeper insight into the advantage of our ASNet, we generate a larger fixation map (28×28) via feeding P into an additional convLSTM with the features from *conv4-3* layer. We therefore derive two baselines: ASNet-14×14 and ASNet-28×28, corresponding to the rough prediction and the refined attention map with a more detailed spatial information.

As shown in Table 2 and Table 3, ASNet-14×14 performs better than previous non-deep learning models and is on par with current top-performing deep learning contenders. Considering our relatively simple network architecture and smaller output resolution (14×14), the suggested ASNet is much favorable and effective. We attribute this primarily to the generalized and powerful saliency representations learned from the SOD task. Additionally, ASNet-28×28 produces further better results, which demonstrates that the proposed ASNet has potential of obtaining better FP results by considering more detailed spatial information.

4.2.2 Performance on the SOD task

Compared SOD models: Here we evaluate the performance of ASNet on its primary task: *SOD*. We perform a quantitative analysis on 3 widely used datasets, namely ECCSD [49], HKU-IS [55] and PASCAL-S [12]. We compare ASNet against 15 recent deep learning based alternatives: Local Estimation and Global Search (LEGS) [59], Multi-scale Deep Feature (MDF) [55], Multi-task Deep Saliency (MDS) [37], Saliency Unified (SU) [41], Deep Contrast Learning (DCL) [57], Encoded Low-level Distance (ELD) [58], Recurrent Fully Convolutional Network (RFCN) [65], Deep Hierarchical Saliency (DHS) [69], Holisitcally-nested Edge Detector based Saliency (HEDS) [66], Non-Local Deep Features (NLDF) [61], Deep Level Sets (DLS) [62], Aggregating MUlti-level convolutional features (AMU) [60], Uncertain Convolutional Features (UCF) [70], Stagewise Refinement Model (SRM) [71], and two-stream Fixation-Semantic CNNs (FSN) [64]. We also consider 4 classical non-deep learning models: Hierarchical Saliency (HS) [49], Discriminative Regional Feature Integration (DRFI) [45], weighted background Contrast (wCtr) [46], and Correspondence-driven Saliency Transfer (CST) [51]. The results are obtained from the authors or by running their public implementations with original settings. Some scores are directly borrowed from literature for making a fair comparison. We also retrain HEDS and AMU on THUS10K [3] and DUT-OMRON [50] datasets to pursue a more fair comparison.

Quantitative SOD results: The precision-recall curves of all methods are plotted in Figure 8. As can be seen, the ASNet outperforms other competitors[1]. We report maximum F-measure and MAE scores in Table 4. Our approach generates the best scores across most datasets. In particular, ASNet shows a significantly improved F-score compared to the second best method, AMU, for the PASCAL-S dataset (0.857 *vs* 0.834), which is one of the most challenging benchmarks. This clearly demonstrates the superior performance of the proposed saliency model in complex scenes. Overall, the proposed ASNet performs favorably against other counterparts over four datasets using all evaluation metrics.

Qualitative SOD results: For an intuitive illustration, we provide the saliency detection results of our ASNet over sev-

1. Here we do not include the results from SU [41], since the authors have not released their code or PR-curve results.

TABLE 5
The F-measure and MAE scores of SOD on four popular datasets: ECCSD [49], HKU-IS [55], PASCAL-S [12], and SOD [77]. ASNet gains the best performance with the assistance of visual attention prior. The best scores are marked in **bold**. See Section 4.2.2 for more details.

Type	Method	ECCSD [49]		HKU-IS [55]		PASCAL-S [12]		SOD [77]	
		F-score ↑	MAE ↓	F-score ↑	MAE ↓	F-score ↑	MAE ↓	F-score ↑	MAE ↓
Non-deep Learning	HS [49]	0.730	0.223	0.710	0.215	0.636	0.259	0.592	0.282
	DRFI [45]	0.787	0.166	0.783	0.143	0.692	0.196	0.712	0.215
	wCtr [46]	0.672	0.178	0.694	0.138	0.611	0.193	0.615	0.213
	CST [51]	0.742	0.147	0.732	0.128	0.598	0.191	0.631	0.210
Deep Learning	MDF [55]	0.831	0.108	0.860*	0.129*	0.764	0.145	0.785	0.155
	LEGS [59]	0.831	0.119	0.812	0.101	0.749	0.155	0.691	0.197
	MDS [37]	0.810	0.160	0.848	0.078	0.818	0.170	0.781	0.150
	DCL [57]	0.898	0.071	0.907	0.048	0.822	0.108	0.832	0.126
	ELD [58]	0.865	0.080	0.844	0.071	0.767	0.121	0.760	0.154
	SU [41]	0.88	0.06	-	-	0.77	0.10	-	-
	RFCN [65]	0.898	0.097	0.895	0.079	0.827	0.118	0.805	0.161
	DHS [69]	0.905	0.061	0.892	0.052	0.820	0.091	0.823	0.127
	HEDS [66]	0.915	0.052	0.913	0.039	0.830	0.080	0.842	0.118
	†HEDS [66]	0.921	0.049	0.916	0.037	0.842	0.076	**0.845**	0.116
	NLDF [61]	0.905	0.063	0.902	0.048	0.831	0.099	0.808	0.126
	DLS [62]	0.825	0.090	0.806	0.072	0.719	0.136	-	-
	AMU [60]	0.889	0.058	0.918	0.052	0.834	0.098	0.773	0.142
	†AMU [60]	0.896	0.055	**0.920**	0.048	0.845	0.087	0.809	0.127
	UCF [70]	0.868	0.068	0.905	0.062	0.771	0.116	0.776	0.148
	SRM [71]	0.910	0.056	0.892	0.046	0.783	0.127	0.792	0.128
	FSN [64]	0.910	0.053	0.895	0.044	0.827	0.095	0.781	0.127
	ASNet	**0.928**	**0.043**	**0.920**	**0.035**	**0.857**	**0.072**	0.835	**0.115**

- The authors in [41] and [62] have not released detailed results or implementations.
* MDF [55] is trained on a subset of HKU-IS, and evaluated on the remaining images.
† indicates the models retained on THUS10K [3] and DUT-OMRON [50] datasets.

eral challenging sample images from above datasets against other 7 state-of-the-art approaches: MDF [55], DCL [57], ELD [58], HEDS [66], DLS [62], UCF [70], and FSN [64]. The qualitative results are depicted in Figure 9. For better visualization, we highlight the main difficulties of each image group. We find that the proposed ASNet is well applicable to various difficult scenarios, such as scenes with multiple objects, cluttered backgrounds and low contrast. Additionally, the saliency values assigned by our ASNet are more confident, compared with other competitors.

4.3 Validation of the Proposed Algorithm

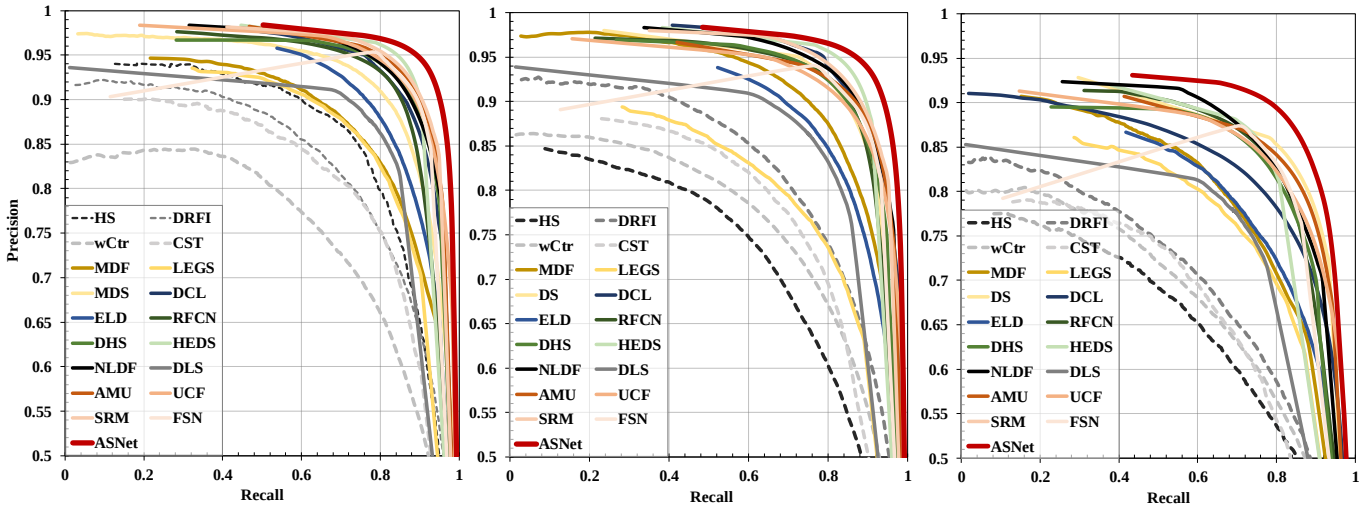
We now conduct a more detailed examination of our proposed approach. We assess 1) contribution of the fixation prior for the SOD task, 2) the effects of convLSTM architecture, 3) the influence of the stacked convLSTMs structure, and 4) the importance of the introduced loss functions.

1. Does fixation prior contribute to SOD? To answer this question, we directly remove the fixation prediction layer and the corresponding loss function $\mathcal{L}_{\text{Att}}$ in Equation 13. After removing the fixation layer, the setting of the next convLSTM is also changed, due to the change of the dimension of input. The whole network can be viewed as a bottom-up/top-down architecture with deep supervision [75]. Then, we retrain ASNet with SOD data and obtain a baseline: *w/o fixation*. From Table 4, we find that fixation map is indeed informative for SOD over all three datasets. The improvement is more pronounced on the PASCAL-S [12] dataset, which is collected from the PASCAL challenge with more general scenes and less center-bias. These results demonstrate that a strong correlation exists between SOD and FP tasks, and our ASNet achieves better performance with the guidance from the fixation map. This also demonstrate that leveraging large-scale FP data could improve the generalization ability of ASNet.

2. What is the effect of convLSTM? Here, we study the contribution of the convLSTM architecture, which constitutes a building block of our ASNet. To this end, we replace the convLSTMs with 5 convolution layers, which have 3×3 kernels and inputs/outputs with original dimensions. Thus, we have a baseline: *w/o convLSTM*. Such network has a similar architecture with previous bottom-up/top-down deep learning models [73], [69]. From Table 4, we observe a drop in F-score and MAE scores over three datasets which implies the effectiveness of the convLSTM. For assessing the effect of the recurrent nature of convLSTM, we further report the performance with different time step settings. It can be observed that convLSTM gradually improves the saliency features step-by-step. We select t=3 for pursuing best performance. A visualization illustration of the iterative saliency improvement via convLSTM can be found in Figure 4. To further study the influence of capacity of convLSTM, we implement ACLNet with different LSTMs with various channel sizes (*i.e.*, 16, 32, 128, 256, with 3 time steps). As demonstrated in Table 4, with the increase of the channel size (16→32→64), better performance is generally achieved. However, further increase of the channel size (64→128→256) hurts the final performance. One reason may be that too many parameters will lead to the model overfitting. In our implementation, we set the channel size as 64, which achieves a relatively better trade-off of performance and parameter amount.

3. Is the hierarchical architecture meaningful? We also study the effect of our hierarchical architecture with a stack of several convLSTMs and top-down saliency inference. We

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TPAMI.2019.2905607, IEEE Transactions on Pattern Analysis and Machine Intelligence



(a) ECCSD (b) HKU-IS (c) PASCAL-S

Fig. 8. SOD results with PR-curve over three widely used benchmarks: ECCSD [49], HKU-IS [55] and PASCAL-S [12], where the scores from non-deep learning models are indicated by dashed lines. Best viewed in color.

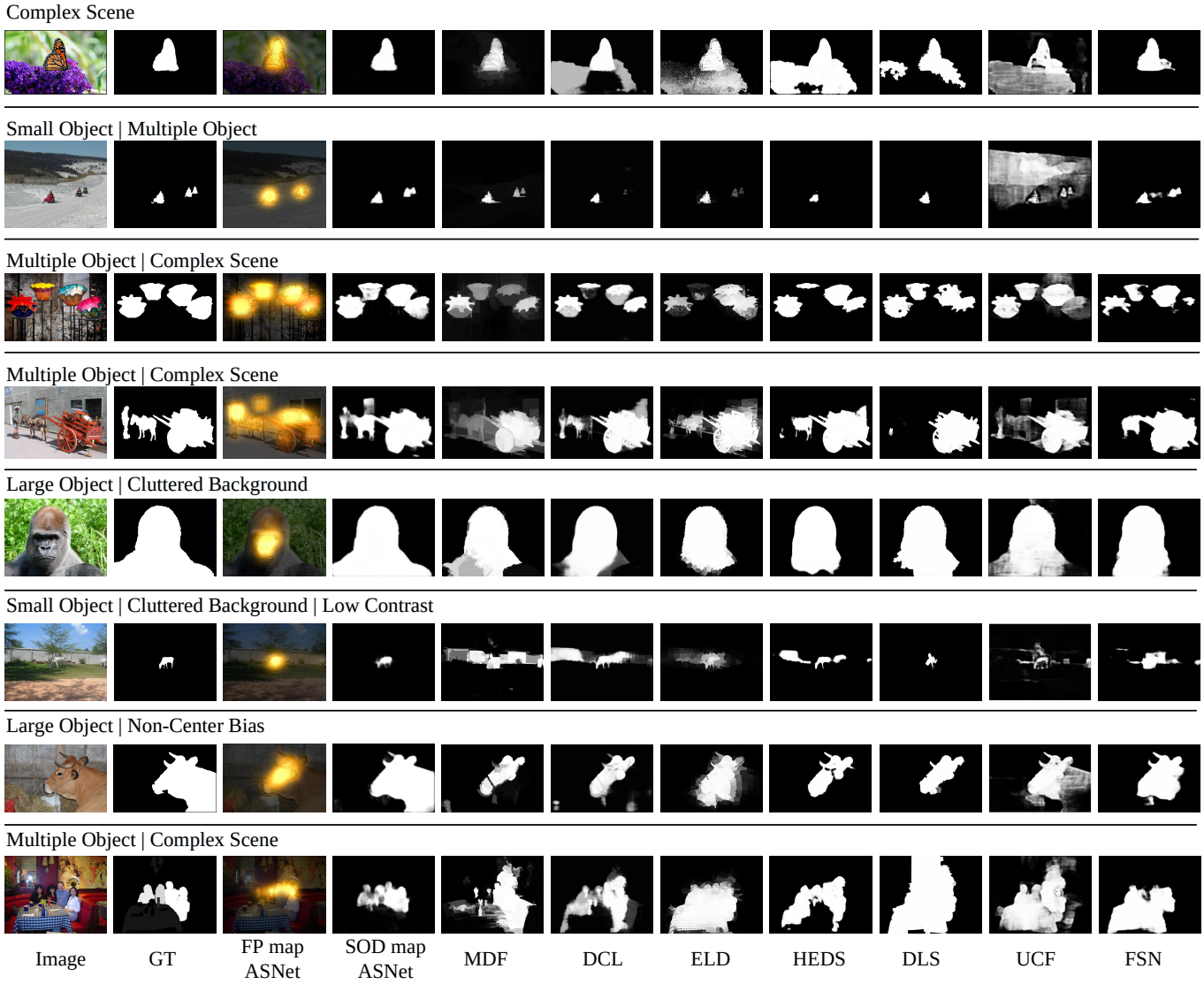


Fig. 9. Qualitative results of the proposed ASNet and other 7 representative SOD models (MDF [55], DCL [57], ELD [58], HEDS [66], DLS [62], UCF [70], and FSN [64]) on sample images. For each example image, we highlight the main challenges and features. For our ASNet, we show both FP and SOD results. It can be observed that ASNet is able to infer object-level saliency maps with the guidance of visual attention predictions. Please see Section 4.2.2 for more details.

TABLE 6
Ablation study of ASNet. We change one component at a time to assess individual contributions. The best scores are marked in **bold**. See Section 4.3 for details.

Aspect	Method	ECCSD [49]		HKU-IS [55]		PASCAL-S [12]	
		F-score ↑	MAE ↓	F-score ↑	MAE ↓	F-score ↑	MAE ↓
reference	ASNet (conv1-output, step $t = 3$, channel size = 64)	**0.928**	0.043	0.920	**0.035**	0.857	**0.072**
variants	w/o fixation	0.913	0.051	0.915	0.040	0.831	0.083
	w/o convLSTM	0.891	0.068	0.887	0.069	0.797	0.112
convLSTM	step $t = 1$, channel size = 64	0.901	0.063	0.892	0.070	0.803	0.107
	step $t = 2$, channel size = 64	0.916	0.051	0.907	0.041	0.842	0.079
	step $t = 4$, channel size = 64	0.926	**0.042**	0.918	0.038	**0.859**	0.073
	step $t = 3$, channel size = 16	0.914	0.049	0.907	0.043	0.846	0.080
	step $t = 3$, channel size = 32	0.925	**0.042**	0.915	0.038	0.855	**0.072**
	step $t = 3$, channel size = 128	0.926	0.045	**0.922**	0.036	0.842	0.074
	step $t = 3$, channel size = 256	0.919	0.048	0.917	0.040	0.848	0.080
architecture	conv5-output	0.853	0.093	0.830	0.079	0.739	0.117
	conv4-output	0.875	0.076	0.844	0.058	0.749	0.092
	conv3-output	0.903	0.061	0.892	0.049	0.794	0.086
	conv2-output	0.919	0.049	0.912	0.040	0.847	0.078
extra loss	w/o $\mathcal{L}_{\mathrm{P}}$	0.923	0.045	0.917	0.038	0.852	0.075
	w/o $\mathcal{L}_{\mathrm{R}}$	0.924	0.046	0.915	0.039	0.854	0.074
	w/o $\mathcal{L}_{\mathrm{F}}$	0.924	0.047	0.916	0.040	0.854	0.074
	w/o $\mathcal{L}_{\mathrm{MAE}}$	0.921	0.044	0.914	0.037	0.850	**0.072**
	w/o extra loss	0.917	0.048	0.912	0.040	0.847	0.075

test 4 baselines: *conv5-output*, *conv4-output*, *conv3-output*, and *conv2-output*, which correspond to the outputs from the intermediate layers of ASNet. Note that the final prediction of ASNEt can be viewed as the output from $conv1$ layer. We find that the saliency results are gradually optimized by adding more details from lower layers.

4. Are the extra loss functions necessary? ASNet is equipped with 4 extra loss functions: $\mathcal{L}_{\mathrm{P}}$, $\mathcal{L}_{\mathrm{R}}$, $\mathcal{L}_{\mathrm{F}}$, $\mathcal{L}_{\mathrm{MAE}}$, which are derived from four widely used SOD metrics. For testing their effects, we retrain ASNet with different loss functions separately and thus we have four baselines: $w/o\ \mathcal{L}_{\mathrm{P}}$, $w/o\ \mathcal{L}_{\mathrm{R}}$, $w/o\ \mathcal{L}_{\mathrm{F}}$, and $w/o\ \mathcal{L}_{\mathrm{MAE}}$. Another baseline $w/o\ extra\ loss$ indicates the results without considering all the extra loss functions. We show their F-measure and MAE scores in Table 5. We observe that these loss functions boost the final performance with about 1% improvement in F-score.

5 Conclusion and Future Work

We proposed a deep learning network, ASNet, towards a better interpretable and efficient SOD model, which leverages eye movements as an indicator for detecting salient objects. The fixation map, as high-level knowledge of a scene, was learned from upper layers of ASNet. Such prior was further utilized for teaching the network where the salient object is and the detailed object saliency was rendered step by step by considering finer and finer features in a top-down manner. ConvLSTM was equipped for iteratively dropping useless features and enhancing the features for better representation. A set of loss functions derived from SOD metrics were introduced for learning more representative features from multiple perspectives and thus further boosting model predictions. Extensive experimental results demonstrate that our approach outperforms several state of the art saliency methods and confirm our view that fixation map is valuable and indicative for SOD. Furthermore, it also works very efficiently, with a fast processing speed of 15 fps.

Our work points out two potential directions of following works. The first one regards exploring the rationale behind SOD from the fixation prediction viewpoint. The second one is to seek better loss functions for boosting the performance of deep learning based SOD models.

References

[1] W. Wang, J. Shen, X. Dong, and A. Borji, "Salient object detection driven by fixation prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 1171–1720.

[2] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

[3] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[4] W. Wang, J. Shen, R. Yang, and F. Porikli, "Saliency-aware video object segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 1, pp. 20–33, 2018.

[5] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Proceedings of the International Conference on Machine Learning*, 2015, pp. 597–606.

[6] B. Alexe, T. Deselaers, and V. Ferrari, "Measuring the objectness of image windows," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 11, pp. 2189–2202, 2012.

[7] R. Zhao, W. Oyang, and X. Wang, "Person re-identification by saliency learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 2, pp. 356–370, 2017.

[8] W. Wang, J. Shen, Y. Yu, and K.-L. Ma, "Stereoscopic thumbnail creation via efficient stereo saliency detection," *IEEE Transactions on Visualization and Computer Graphics*, vol. 23, no. 8, pp. 2014–2027, 2017.

[9] W. Wang, J. Shen, and H. Ling, "A deep network solution for attention and aesthetics aware photo cropping," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[10] A. Borji, "What is a salient object? A dataset and a baseline model for salient object detection," *IEEE Transactions on Image Processing*, vol. 24, no. 2, pp. 742–756, 2015.

[11] A. Borji, D. N. Sihite, and L. Itti, "What stands out in a scene? A study of human explicit saliency judgment," *Vision Research*, vol. 91, pp. 62–77, 2013.

[12] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280–287.

[13] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 5706–5722, 2015.

[14] X. Shi, N. D. Bruce, and J. K. Tsotsos, "Fast, recurrent, attentional modulation improves saliency representation and scene recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2011 IEEE Computer Society Conference on*, 2011, pp. 1–8.

[15] B. J. White, D. J. Berg, J. Y. Kan, R. A. Marino, L. Itti, and D. P. Munoz, "Superior colliculus neurons encode a visual saliency map during free viewing of natural dynamic video," *Nature communications*, vol. 8, p. 14263, 2017.

[16] J. Tsotsos, I. Kotseruba, and C. Wloka, "A focus on selection for fixation," *Journal of Eye Movement Research*, vol. 9, no. 5, 2016.

[17] S. Xingjian, Z. Chen, H. Wang, D.-Y. Yeung, W.-K. Wong, and W.-C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," in *Advances in Neural Information Processing Systems*, 2015, pp. 802–810.

[18] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial Intelligence*, vol. 78, no. 1-2, pp. 507–545, 1995.

[19] 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.

[20] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*, 2007, pp. 545–552.

[21] J. Zhang and S. Sclaroff, "Saliency detection: A boolean map approach," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 153–160.

[22] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 10, pp. 1915–1926, 2012.

[23] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 185–207, 2013.

[24] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, no. 5, pp. 802–817, 2006.

[25] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, no. 7, pp. 32–32, 2008.

[26] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *Advances in Neural Information Processing Systems*, 2005, pp. 481–488.

[27] N. D. Bruce and J. K. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *Journal of Vision*, vol. 9, no. 3, pp. 5–5, 2009.

[28] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007, pp. 1–8.

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

[30] A. Borji, "Saliency prediction in the deep learning era: An empirical investigation," *arXiv preprint arXiv:1810.03716*, 2018.

[31] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2798–2805.

[32] S. S. Kruthiventi, K. Ayush, and R. V. Babu, "Deepfix: A fully convolutional neural network for predicting human eye fixations," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4446–4456, 2017.

[33] X. Huang, C. Shen, X. Boix, and Q. Zhao, "SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 262–270.

[34] N. Liu, J. Han, T. Liu, and X. Li, "Learning to predict eye fixations via multiresolution convolutional neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 2, pp. 392–404, 2018.

[35] J. Pan, E. Sayrol, X. Giro-i Nieto, K. McGuinness, and N. E. O'Connor, "Shallow and deep convolutional networks for saliency prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 598–606.

[36] W. Wang and J. Shen, "Deep visual attention prediction," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2368–2378, 2018.

[37] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3919 – 3930, 2016.

[38] W. Wang, J. Shen, F. Guo, M.-M. Cheng, and A. Borji, "Revisiting video saliency: A large-scale benchmark and a new model," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.

[39] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations*, 2015.

[40] S. Jetley, N. Murray, and E. Vig, "End-to-end saliency mapping via probability distribution prediction," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5753–5761.

[41] S. S. Kruthiventi, V. Gudisa, J. H. Dholakiya, and R. Venkatesh Babu, "Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5781–5790.

[42] W. Wang, J. Shen, F. Porikli, and R. Yang, "Semi-supervised video object segmentation with super-trajectories," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[43] W. Wang, J. Shen, X. Li, and F. Porikli, "Robust video object cosegmentation," *IEEE Transactions on Image Processing*, vol. 24, no. 10, pp. 31–37, 2015.

[44] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1597–1604.

[45] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2083–2090.

[46] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2814–2821.

[47] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou, "Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation," *IEEE Transactions on Image Processing*, vol. 27, no. 2, pp. 568–579, 2018.

[48] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," *European Conference on Computer Vision*, pp. 29–42, 2012.

[49] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1155–1162.

[50] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173.

[51] W. Wang, J. Shen, L. Shao, and F. Porikli, "Correspondence driven saliency transfer," *IEEE Transactions on Image Processing*, vol. 25, no. 11, pp. 5025–5034, 2016.

[52] W. Wang, J. Shen, and L. Shao, "Consistent video saliency using local gradient flow optimization and global refinement," *IEEE Transactions on Image Processing*, vol. 24, no. 11, pp. 4185–4196, 2015.

[53] W. Wang, J. Shen, H. Sun, and L. Shao, "Video co-saliency guided co-segmentation," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 8, pp. 1727–1736, 2018.

[54] F. Guo, W. Wang, J. Shen, L. Shao, J. Yang, D. Tao, and Y. Y. Tang, "Video saliency detection using object proposals," *IEEE Transactions on Cybernetics*, 2017.

[55] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5455–5463.

[56] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1265–1274.

[57] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 478–487.

[58] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 660–668.
[59] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3183–3192.
[60] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 202–211.
[61] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6593–6601.
[62] P. Hu, B. Shuai, J. Liu, and G. Wang, "Deep level sets for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 540–549.
[63] W. Wang, J. Shen, and L. Shao, "Video salient object detection via fully convolutional networks," *IEEE Transactions on Image Processing*, vol. 27, no. 1, pp. 38–49, 2018.
[64] X. Chen, A. Zheng, J. Li, and F. Lu, "Look, perceive and segment: Finding the salient objects in images via two-stream fixation-semantic CNNs," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1050–1058.
[65] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *European Conference on Computer Vision*, 2016, pp. 825–841.
[66] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3203–3212.
[67] M. Amirul Islam, M. Kalash, and N. D. B. Bruce, "Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018.
[68] J. Kuen, Z. Wang, and G. Wang, "Recurrent attentional networks for saliency detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3668–3677.
[69] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 678–686.
[70] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 212–221.
[71] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 4039–4048.
[72] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam, "Pyramid dilated deeper convLSTM for video salient object detection," in *The European Conference on Computer Vision*, 2018.
[73] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *European Conference on Computer Vision*, 2016, pp. 75–91.
[74] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.
[75] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *AISTATS*, 2015, pp. 562–570.
[76] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1072–1080.
[77] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition - Workshops*, 2010.
[78] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *International Conference on Learning Representations*, 2015.
[79] M. Everingham, L. V. Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes (voc) challenge," *International Journal of Computer Vision*, vol. 88, no. 2, pp. 303–338, 2010.
[80] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2001, pp. 416–423.
[81] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics," in *Proceedings of the IEEE International Conference on Computer Vision*, 2014, pp. 1153–1160.
[82] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vision Research*, vol. 45, no. 18, pp. 2397–2416, 2005.
[83] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," in *MIT Technical Report*, 2012.

Wenguan Wang received his PhD degree from Beijing Institute of Technology in 2018. He is currently a research scientist at Inception Institute of Artificial Intelligence, UAE. From 2016 to 2018, he was a joint Ph.D. candidate in Department of Statistics, University of California, directed by Prof. Song-Chun Zhu. He received the Baidu Scholarship in 2016. His current research interests include computer vision, image processing and deep learning.

Jianbing Shen (M'11-SM'12) is a Professor with the School of Computer Science, Beijing Institute of Technology. He has published about 100 journal and conference papers such as *IEEE TPAMI*, *IEEE CVPR*, and *IEEE ICCV*. He obtained many flagship honors including the Fok Ying Tung Education Foundation from Ministry of Education, the Program for Beijing Excellent Youth Talents from Beijing Municipal Education Commission, and the Program for New Century Excellent Talents from Ministry of Education. He is an Associate Editor of IEEE TNNLS and Neurocomputing.

Xingping Dong received the B.S. degree in information and computing science and the second B.S. degree in computer science and technology from Xiamen University. He is currently working toward the Ph.D. degree in the School of Computer Science, Beijing Institute of Technology, Beijing, China. His current research interests include object saliency, deep reinforcement learning, and visual object tracking.

Ali Borji received the PhD degree in cognitive neurosciences from the Institute for Studies in Fundamental Sciences (IPM), 2009. He is currently an assistant professor at Center for Research in Computer Vision, University of Central Florida. His research interests include visual attention, visual search, machine learning, neurosciences, and biologically plausible vision models.

Ruigang Yang received the MS degree from Columbia University in 1998 and the PhD degree from the University of North Carolina, Chapel Hill in 2003. He is currently a full professor of Computer Science at the University of Kentucky. His research interests span over computer vision and computer graphics, in particular in 3D reconstruction and 3D data analysis. He has published more than 100 papers, which, according to Google Scholar, has received close to 6,000 citations with an h-index of 37 (as of 2014). He is currently an associate editor of the *IEEE Trans. on Pattern Analysis and Machine Intelligence* and a senior member of IEEE.