# Bottom–Up Visual Saliency Estimation With Deep Autoencoder-Based Sparse Reconstruction

Chen Xia, Fei Qi, *Member, IEEE*, and Guangming Shi, *Senior Member, IEEE*

***Abstract*— Research on visual perception indicates that the human visual system is sensitive to center–surround (C–S) contrast in the bottom–up saliency-driven attention process. Different from the traditional contrast computation of feature difference, models based on reconstruction have emerged to estimate saliency by starting from original images themselves instead of seeking for certain *ad hoc* features. However, in the existing reconstruction-based methods, the reconstruction parameters of each area are calculated independently without taking their global correlation into account. In this paper, inspired by the powerful feature learning and data reconstruction ability of deep autoencoders, we construct a deep C–S inference network and train it with the data sampled randomly from the entire image to obtain a unified reconstruction pattern for the current image. In this way, global competition in sampling and learning processes can be integrated into the nonlocal reconstruction and saliency estimation of each pixel, which can achieve better detection results than the models with separate consideration on local and global rarity. Moreover, by learning from the current scene, the proposed model can achieve the feature extraction and interaction simultaneously in an adaptive way, which can form a better generalization ability to handle more types of stimuli. Experimental results show that in accordance with different inputs, the network can learn distinct basic features for saliency modeling in its code layer. Furthermore, in a comprehensive evaluation on several benchmark data sets, the proposed method can outperform the existing state-of-the-art algorithms.**

***Index Terms*— Autoencoder, center–surround (C–S) difference, deep learning, reconstruction, saliency, unsupervised feature learning.**

## I. Introduction

EMBRACED by enormous amounts of incoming information from the visual environment, the human visual system (HVS) can cope with the situation effortlessly with limited computational resources and capacity [1], [2]. One main mechanism to reduce the complexity of visual processing is the selective attention of guiding the analysis mainly on a small group of parts in the visual field while leaving the rest to only limited processing [2]–[4]. Bottom–up saliency, which is scene-dependent and stimulus-driven, plays an important part in the selection process to achieve the computational efficiency [5]. Modeling the way that humans locate the information of interest rapidly in free-viewing natural scenes has significant impacts on both neuroscience and computer vision. For one thing, it can provide a better insight into the essential principles and underlying mechanisms of the HVS. For another, saliency-inspired methods can also offer solutions to the applications with analogous demands, such as image and video compression [6], [7], object detection [8], [9], and object recognition [10], [11].

Similar to the structure of the retina which distributes different resolutions for the central fovea and peripheral area, Itti *et al.*'s [3] classical cognitive center–surround (C–S) bottom–up saliency model was proposed to estimate saliency by the across-scale subtraction on the feature maps of color, luminance, and orientation. Based on their work, C–S contrast has become one of the most influential mechanisms in low-level saliency. To solve the problem of C–S contrast measurement, many fixation prediction algorithms have been proposed [12]–[14]. However, for the models built on local image-processing techniques, it is often hard to provide satisfactory results on the images with texture structures [15]. In view of this, Wu *et al.* [16] began to extend the C–S model to nonlocal regions and proposed a redundancy reduction-based approach [16]. In [15] and [17], to further integrate the influence of distinct surrounding patches on the central area in a unified and optimized manner, saliency was estimated by the reconstruction residual of representing the central patch with a linear combination of its surrounding patches in a nonlocal region. Although the C–S comparison scheme of reconstruction can exhibit a better ability to detect rarity and saliency than feature distinction, without sufficient consideration of global competition, the algorithms may fail when one location has low local and nonlocal C–S difference, but it is still unusual in the whole visual scene.

Global rarity also plays an important role in saliency estimation [18]. To model global saliency, recent methods measure the global contrast of each region based on spatially weighted feature dissimilarities [19]–[21] or highlight the global rarity of features from the aspect of probability distribution [18], [22]–[24]. In spite of the importance of local C–S contrast and global rarity, the separate consideration of either of the two cues to estimate saliency is not sufficient.

Manuscript received September 29, 2014; revised December 13, 2015; accepted December 20, 2015. This work was supported in part by the Major State Basic Research Development Program of China (973 Program) within the Ministry of Science and Technology, China, under Grant 2013CB329402, in part by the National Natural Science Foundation of China under Grant 61572387, Grant 61227004, and Grant 61472301, in part by the Ministry of Education, China, under Grant 20130203130001, in part by the International Cooperation Project of Shaanxi Science and Technology Research and Development Program under Grant 2014KW01-02, and in part by the Fundamental Scientific Research Funds of Xidian University under Grant K5051302012.

The authors are with the Key Laboratory of Intelligent Perception and Image Understanding, School of Electronic Engineering, Ministry of Education, Xidian University, Xi'an 710071, China (e-mail: cxia@stu.xidian.edu.cn; fred.qi@ieee.org; gmshi@xidian.edu.cn).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2015.2512898

Therefore, models have combined their results to benefit from the advantages of the two parts and achieved a better performance than the approaches with an emphasis on one-side component [17], [18]. However, existing approaches usually treat local and global saliency as individual channels without considering the complementarity between them.

Inspired by the reconstruction-based saliency detection approaches [15], [17], [25], [26] and successful application of deep autoencoder networks [27]–[29], we construct a deep multilayer architecture to explore a novel adaptive C–S comparison scheme by incorporating global competition into a framework of nonlocal C–S reconstruction and comparison. In the proposed method, the deep architecture is constructed by connecting a traditional autoencoder with an inference layer. The network is trained by global sampling data to reconstruct (predict) the central patch from its surrounding patch. Based on the learned C–S inference pattern, the saliency of each position is estimated according to the residual between the reconstructed and original central patches.

Compared with existing saliency estimation methods, the proposed one presents several attractive characteristics. First, the global data of the current image can directly influence the computation of C–S inference. In other words, the measure of the C–S relationship of each region is not individual but fused with global comparison, which can further distinguish the unusual salient regions from the background if they share similar nonlocal C–S distinction. In this way, the proposed method can achieve a better performance than the models focusing merely on nonlocal C–S contrast and the ones treating global rarity and local C–S contrast individually [17], [18]. In addition, in contrast to the postprocessing fusion way of global saliency, integrating global rarity in the initial competition is consistent with the acquisition of contextual information in the early stage of visual-processing chain [30].

Second, the proposed model can adaptively optimize the selection and competition of features. Among the previous models of saliency, a tremendous amount of research has been made on feature selection and fusion [22], [31]–[33]. With the increasing dimensionality of features, how to assign weights to multiple feature maps has also become a challenging problem. To solve it, a common way is to use equal weights for all the visual features [3], [13] or to derive their contributions by supervised training [31], [34]. However, it is suggested that the importance and contributions of features should vary with visual fields [35]. Fixed weights of hand-crafted features may not be suitable for every input. In this paper, by learning from the current scene, the proposed model can achieve the feature extraction and interaction simultaneously in an adaptive way, which can form a better generalization ability to handle various stimuli. Based on the traditional idea of estimating saliency with constant features and C–S contrast inference mechanism, this paper can offer a different direction to understand and model visual saliency.

Finally, the learning procedure of the proposed model is independent of labeled data. In the previous studies, most of the learning-based saliency models are based on a set of labeled training samples. However, using supervised techniques for saliency detection may have several challenges, such as the acquisition of fixation data [36]. Furthermore, with the prevalence of deep learning, the research has focused more on exploring from the unlabeled data to answer the question what unlabeled data can tell us [27]–[29]. In this paper, starting from original image information rather than the ground truth (GT) of human fixations, the model can provide a solution to the stimulus-driven saliency with the network of powerful ability to learn from the unlabeled data. When the labeled data are scarce or even nonexistent, our saliency detection method can be considered as a complement to the traditional supervised learning-based approaches [31], [34], [37]–[40].

The rest of this paper is organized as follows. Section II reviews related works. Section III introduces the proposed saliency estimation method using a deep autoencoder architecture. In Section IV, we compare the performance of the proposed model with the existing state-of-the-art saliency estimation methods. The discussions and conclusions are given in Sections V and VI, respectively.

## II. Related Works

As a concept broader than saliency, visual attention is dictated by two mechanisms: 1) the bottom–up data-driven mechanism and 2) the top–down task-driven mechanism [41]. In contrast to the top–down mechanism which is influenced by more high-level cognitive factors, such as semantic understanding, memories, and tasks, saliency, on the other hand, refers mainly to the bottom–up and stimulus-driven process.

To model visual saliency, as one seminal work, Itti *et al.* [3] performed C–S operations on a set of early visual features with the core hypothesis that the saliency of the current location is determined by its degree of distinction to the surroundings. Following the C–S hypothesis, during the past decade, various measuring strategies of C–S contrast have been presented to model saliency. Ma and Zhang [12] defined the colors of an image as a perceive field with each pixel as a perception unit, and measured the saliency of a central pixel by its distinction to other units in the surrounding region. To take the immediate context rather than an isolated pixel into consideration, direct feature dissimilarities between patches were utilized to represent the local and nonlocal uniqueness [16], [18]. Similarly, Seo and Milanfar [32] measured the matrix cosine similarities of the central feature matrix to all the surrounding ones to represent saliency. Different from the patchwise feature difference, other methods aggregated the influence of distinct surrounding regions on the center in a unified manner. Gao and Vasconcelos [13] defined salient areas as the locations with the most discriminant power to classify the center and surround according to the feature responses. Klein and Frintrop [14] used the Kullback–Leibler divergence (KLD) to estimate the C–S difference of feature statistics and combined the conspicuity maps of multiple features into a saliency map.

It has been suggested that in the perceptual procedure, sensory cortex may form an ability to predict the input and to suppress the response to the expected part of the field [1], [42]. In recent years, to simulate the prediction process and to emphasize the unexpected regions, reconstruction-based

approaches have been proposed from a new perspective to compute C–S discrepancy with outstanding performance. Beginning with the original image itself instead of seeking for certain features, Xia *et al.* [15], [17] represented the central patch with its surrounding patches in a nonlocal region and measured saliency with the sparse reconstruction residual. To solve the problem of spatiotemporal saliency detection, Ren *et al.* [25] modeled both the temporal and spatial saliency based on a regularized feature reconstruction framework and combined their results to implement the final saliency estimation. With the background templates extracted from the boundaries of each image, Li *et al.* [26] reconstructed each image segment by the bases from the corresponding background regions and integrated dense and sparse reconstruction errors to compute saliency.

Besides the exploration of ways to compute local and nonlocal C–S contrast, some models have measured saliency from the aspect of global rarity. For instance, Hou and Zhang [43] found the anomalous regions of each image by spectral residual with the log spectrum computed from the entire image. Cheng *et al.* [19] proposed the models of global contrast to accumulate the spatially weighted dissimilarities between the current region and others from the whole image [21] or a set of similar images [44]. Instead of using all the image patches for comparison, in [20], $K$ most similar patches were extracted for each area to calculate the global uniqueness. In addition, the distribution of features has also been adopted to model global saliency. Bruce and Tsotsos [22] computed the self-information of features for each region with respect to the global surrounding region. Hou and Zhang [23] introduced the incremental coding length to measure the entropy gain of each feature with the distribution of features sampled from the entire scene. Zhang *et al.* [24] proposed a definition of bottom–up saliency similar to [22], with the main difference that they used the statistics of feature distribution from a training set rather than the current image. In recent years, to combine the advantages of local and global saliency, methods have begun to aggregate their maps into one model, and more research in this direction emerged. Borji and Itti [18] modeled global saliency as the inverse of the probability of each patch over the image to fuse with the local rarity based on weighted feature dissimilarity. Similarly, Xia *et al.* [17] utilized the global self-information of color to provide a complement to the nonlocal saliency of reconstruction residual and achieved a better performance than the model with separate consideration.

To pop out salient regions, in the previous studies, efforts have also been made on feature selection. While a majority of computational models [3], [13], [14], [24] used the traditional features of intensity, color, and orientation proposed by feature integration theory [45], some methods were aimed at finding certain particular features such as symmetry [33], gist [30], and local steering kernel [32] for directing saliency. In the past few years, with the gaining momentum of feature learning in computer vision community, some other algorithms have predicted saliency with the features obtained by learning rather than the hand-crafted image features. Bruce and Tsotsos [22] first performed independent component analysis on a large sample of natural image patches to learn V1-like features. Similarly, Borji and Itti [18] learned a dictionary of basis functions from a repository of natural scenes. The features for rarity calculation were then derived by projecting image patches into the space of the dictionary [18]. Based on the high-dimensional features extracted by different visual properties, Shen and Wu [46] trained a linear transformation on the original feature space to further separate salient regions from the background. More recently, Vig *et al.* [38] searched for the optimal combinations of multilayer features from a large pool of hierarchical neuromorphic models to take advantage of the diversity of features with various visual representations.

Different features can highlight saliency from distinct aspects. To model saliency under more diversified scenes, recent improvements have been achieved mainly by adding more and more features [31], [34], [37], [47]. With the features of increasing dimensions and types, machine learning has also become a popular tool to combine the contributions of distinct features in an optimized way instead of aggregating them with equal or designed weights [8], [31], [34], [37], [39], [40], [47], [48]. A review of learning-based saliency detection can be found from [36]. Despite the outstanding performance of learning-based methods, the majority of them are dependent on a large number of labeled samples to train the saliency predictors of mapping high-dimensional feature vectors into scalar saliency values. Unfortunately, unlike other supervised learning applications in computer vision [49]–[52], obtaining GT for saliency prediction is a less straightforward and labor intensive work due to the normal unavailability of the expensive devices and large-scale eye-tracking experiments [36]. Therefore, how to learn from natural images themselves to mine the underlying relationship between the data and the saliency has become a problem to be resolved, which is also the key concern of bottom–up saliency modeling.

With powerful ability to learn latent representations from the unlabeled data, unsupervised feature learning and deep learning have emerged and presented the state-of-the-art performance in a wide variety of tasks [27]–[29], [49], [52]–[54]. However, it should be noted that most of them concentrated on traditional recognition problems. As for the area of detection, its applications are usually restricted to specific object categories, such as pedestrian [50], face [51], robotic grasp [55], and road in high-resolution aerial images [56]. It was not until recently that deep learning algorithms were successfully utilized to solve the general object detection problems with marked improvements over the best previous results. Based on the convolutional neural networks (CNNs) defined by [49] which showed powerful object representation ability and outstanding performance on a challenging large-scale classification task, Szegedy *et al.* [53] formulated object detection as a CNN-based regression problem to generate object bounding box masks in a multiscale fashion. Girshick *et al.* [54] also took advantage of the CNN described in [49] to extract features from the region proposals and optimized one linear support vector machine (SVM) per class with the learned features and training labels, which can achieve much better results than the regression-based model [53] in practice. To scale up to multiple

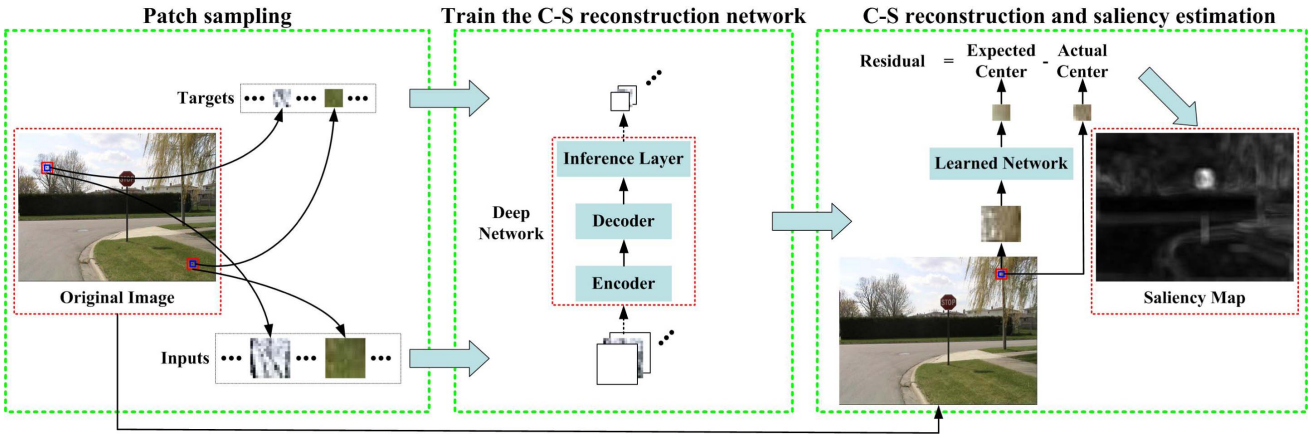


Fig. 1. Diagram of the proposed saliency estimation algorithm. First, pairs of surrounding (red patches) and central patches (blue patches) are randomly sampled from the original image according to a uniform distribution. Then, the C–S inference network is trained with the sampled data as inputs and corresponding targets. Finally, the learned network is applied on each region of the input image to predict the center and compare it with the actual center to estimate unpredictability and saliency.

classes' detection, unlike the previous methods of training detectors separately for each class, Erhan *et al.* [9] proposed a saliency-inspired class-agnostic object detection model by solving an assignment problem to predict the bounding boxes of potential objects and the confidence scores of how likely they contain objects.

While deep models have begun to be leveraged in the research on object detection, regarding it as methodologies for saliency estimation still remains challenging. Based on the findings that deep learning models share some similar properties with the early processing stage of the primate visual system [57], Shen *et al.* [39] and Shen and Zhao [40] used a multilayer network to learn mid-level and high-level features from collected salient regions and integrated the learned features with a linear SVM as [31] to highlight the conspicuity of semantic object regions. Although the sparse coding algorithm [58], they used to generate hierarchical features from input data is an unsupervised scheme, the inputs themselves are derived in an uneven sampling manner based on labeled information. With the third layer responses learned from salient regions as features, their model is more suitable for the images with high-level concepts [31], [59]. As for the data sets, including more general natural scenes [22], [33], the advantage of their model declines. In this paper, to achieve consistent performance over all the data sets, we lay more emphasis on the bottom–up saliency estimation to explore the intrinsic attributes and principles of the HVS shared by free-viewing different scenarios.

## III. Saliency Estimation With Autoencoder-Based Reconstruction

Motivated by the reconstruction-based saliency and powerful representation of autoencoders with nonlinear encoder and decoder functions, we propose a deep autoencoder-based reconstruction network (AER) for visual saliency estimation. It enables us to deal with the input data lying on a curved manifold rather than a linear manifold in a high-dimensional space and, therefore, can achieve a more general representation of C–S discrepancy under diversified conditions. An overview of our proposed framework for saliency estimation is shown in Fig. 1. In this section, we first introduce the deep network to estimate the C–S relationship and visual saliency. Then, we provide the detailed description of the network training process. Finally, we present how to apply the learned inference network to compute the reconstruction-based saliency.

### A. C–S Inference Network

In recent years, deep learning has shown remarkable ability to discover a compact representation (or features) from the data in many applications [28], [39], [40], [49], [54], [60]. Among different deep network models, autoencoder networks have two properties desirable for bottom–up saliency detection. First, the code vector in the central bottleneck layer can form a compressed representation of input which can present much better reconstruction than other dimension reduction methods, including principal components analysis [27]. Hence, the features for C–S relationship representation and reconstruction can be derived by learning adaptively from the images. Second, because the output of an autoencoder network is the reconstruction of input, different from the feature extraction of CNN [52], [54], features can be built directly from the unlabeled data. It means that in this paper, the saliency estimation of each image can merely be related to the visual scene itself without supervised information out of the current scene.

The architecture of our deep C–S inference network is shown in Fig. 2. In the virtue of the aforementioned properties of autoencoders, the bottom part of the network is corresponding to a typical autoencoder network [27], [29] to discover a compact and economical representation for the inputs. Furthermore, based on the observation that it is signal contrast rather than absolute strength that guides our bottom–up saliency [4], [12], an extra inference layer is added at the top to provide ways to explore the C–S contrast relationship. By training the deep network of asymmetric architecture, the saliency model can better simulate the retina with general

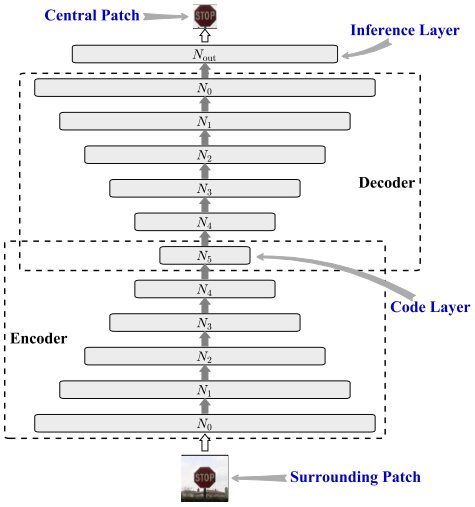


Fig. 2. Architecture of the deep C–S inference model. The gray up arrows denote the connection of the deep network and the white up arrows denote the transformation of data format. The former two parts of the network, encoder and decoder, compose a typical autoencoder network, while an extra inference layer is added at the top to provide ways to explore the C–S contrast relationship.

principles of being sensitive to the C–S discontinuities and processing information unevenly for the fovea and peripheral regions.

To the best of our knowledge, there is almost no theory to guide the specific selection of the network architecture. However, we found in the experiments that a simple way of progressively decreasing or halving the number of units in each layer as [27] and [29] has already worked. Therefore, the sizes of layers in the encoder generally follow $N_i \geq 2N_{i+1}$ for $i = 1, \ldots, 4$, with $N_0$ being determined by the dimensionality of the surrounding patch. The decoder shares these settings with a symmetric structure. In addition, similar to [27], all the units are logistic except for the $N_5$ linear units in the central code layer. With the bottleneck layer of $N_5$ units, the deep network is forced to discover some internal representation and relations in high-dimensional input data, which will also be validated by visualization in Section IV-D.

### *B. Train the Deep C–S Inference Network*

Given the structure of the network, we are then faced with the question of how to train the network for C–S contrast measurement. First, to construct the training samples feeding into the network and their respective targets, we take $m$ pixels randomly from each image. Then, as shown in Fig. 1, for each chosen pixel, we treat the surrounding region with a larger size ($D \times D$ pixels) as the input and its central part with a smaller size ($d \times d$ pixels) as the corresponding target. The effect of various parameters on the overall saliency detection performance will be discussed in Section IV. With the global sampling step, the patches with different scales can be extracted for competition as they are from distinct depth ranges of the image. Therefore, the issue of scale can be considered implicitly in the proposed method, instead of modeling it in an explicit way.

It should be pointed out that for a given image, the inference network is trained merely with $m$ pairs of samples from the current image itself. Therefore, for the sampling results, foreground regions usually present a smaller probability of being sampled than background regions. On the other hand, the background will have more contribution to the learning process of the network parameters. That is to say, in the competition among samples, the unified C–S inference pattern learned by the model will deviate from the description of C–S connection in salient regions to enhance their reconstruction error and saliency. This idea is analogous to that in [44] where saliency detection is formulated as a sampling problem and salient patches are defined as the regions with the least probability of being sampled from a large corpus of unlabeled images. The main difference is that in our method, the saliency computation of each area does not rely on any information outside of the current image. As a result, our model is more consistent with the modeling of stimulus-driven bottom–up saliency.

Before the training phase, to achieve the contrast normalization among distinct patches as [29], [39], and [56], a preprocessing procedure is applied to the sampled data to transform them from the original red-green-blue color space to the opponent color space with independent channels [61] and to normalize them to the range of [0, 1].

When it comes to training deep autoencoders, it was very difficult to optimize the network parameters by backpropagation learning procedure until the proposition of using unsupervised layer-by-layer pretraining [27], [62]. Therefore, to train our deep C–S inference network, we first learn a stack of restricted Boltzmann machines (RBMs) to initialize the deep autoencoder for the subsequent backpropagation. In the pretraining, the activation probabilities of the hidden units derived by training the current RBM are then taken as the visible units to train the next RBM in the stack. Take the RBM with binary visible and hidden units, for example, we use the standard contrastive divergence [63] to update the weights and biases of the connections with the following steps.

1) Initialize the RBM with zero biases and small random weights chosen from a zero-mean Gaussian with a standard deviation of 0.1.
2) In the positive phase, given the values of the visible units $\boldsymbol{v_1}$, for each hidden unit $j$, extract the binary state $h_{1j} \in \{0, 1\}$ from the conditional probability $p(h_{1j} = 1|\boldsymbol{v_1})$, which is computed by

$$p(h_{1j} = 1|\boldsymbol{v_1}) = \sigma\left(b_j + \sum_i v_{1i} w_{ij}\right) \tag{1}$$

where $b_j$ is the bias of the hidden unit $j$, and $w_{ij}$ is the weight on the connection between the visible unit $i$ and the hidden unit $j$. In addition, $\sigma(x) = 1/(1 + \exp(-x))$ is the logistic function.
3) In the negative phase, inspired by the work in [64], reconstruct visible units $\boldsymbol{v_2}$ with (2) by taking the probability directly instead of stochastically picking a binary value with the probability. The activation probabilities $p(h_{2j} = 1|\boldsymbol{v_2})$ of the hidden units $j$ are calculated by

replacing $\boldsymbol{v_1}$ with $\boldsymbol{v_2}$ in (1)

$$v_{2i} = p(v_{2i} = 1|\boldsymbol{h}_1) = \sigma\left(a_i + \sum_j h_{1j} w_{ij}\right) \quad (2)$$

where $a_i$ is the bias of the visible unit $i$.

4) Update the weights and biases by the following rules:

$$\begin{aligned} \triangle \boldsymbol{W} &\propto \boldsymbol{v}_1(p(\boldsymbol{h}_1 = 1|\boldsymbol{v}_1))^T - \boldsymbol{v}_2(p(\boldsymbol{h}_2 = 1|\boldsymbol{v}_2))^T \\ \triangle \boldsymbol{a} &\propto \boldsymbol{v}_1 - \boldsymbol{v}_2 \\ \triangle \boldsymbol{b} &\propto p(\boldsymbol{h}_1 = 1|\boldsymbol{v}_1) - p(\boldsymbol{h}_2 = 1|\boldsymbol{v}_2). \end{aligned} \quad (3)$$

For the architecture in Fig. 2, after training the five-layer RBMs to initialize the weights for the encoder, we then take the transposes of those weights to initialize the decoding part. In addition, we give random weights to the top inference layer in the same way as the initialization of the RBMs.

In the backpropagation process, based on the initial parameters, we fine-tune the whole network globally by minimizing the cross-entropy error between outputs and targets

$$E(\boldsymbol{W}, \boldsymbol{a}, \boldsymbol{b}) = -\sum_i (t_i \log o_i + (1 - t_i) \log(1 - o_i)) \quad (4)$$

to establish a bridge between the surround and the center and to derive a novel C–S contrast comparison scheme, where $t_i$ and $o_i$ are the values of element $i$ in the target vector (vector of actual center) and output vector (vector of reconstruction center), respectively. The total error to be optimized is the average of the errors over all the training cases.

Despite the insensitivity to the learning parameters of the proposed model, it is necessary to present them for reference. To train the RBMs, we use the minibatches of 100 with 100 epochs, while in the backpropagation process, we use the minibatches of 200 with ten epochs. For the first four layers of the RBMs, the learning rate is taken as 0.1 for all the parameters and for the fifth RBM with the real-valued feature detectors drawn from a unit variance Gaussian, we use the learning rate of 0.001. Moreover, to raise the speed of learning, momentum is utilized in the learning process of the RBMs with an initial value of 0.5 and increased to 0.9 after five epochs.

### C. Reconstruction-Based Saliency Estimation

After the training process, we have obtained a unified network to predict the center with the surround as input. Despite identical reconstruction parameters, the different regions of an image present distinct reconstruction ability due to the competition in sampling and learning phases. As shown in Fig. 3, because of the randomly initialized parameters of the top layer, the reconstructions of the pretraining stage cannot convey any effective content of the actual targets. After the backpropagation procedure by training with pairs of surrounding and central patches, most of the background patches have learned to capture the information about the color and structure of the targets. Nevertheless, the trained network still cannot model the C–S relationship in salient regions well because of fewer foreground samples. As a result,

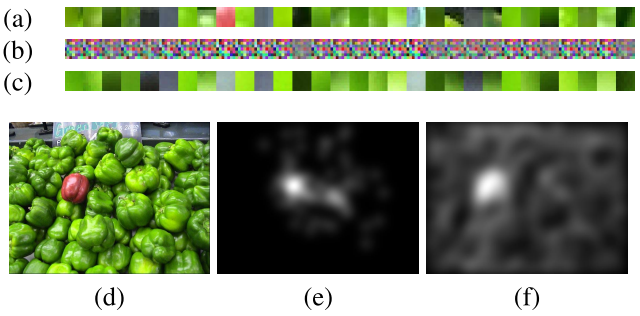


Fig. 3. Reconstructions of the center by surrounding patches. (a) Targets of the central patches. (b) Initial inference results of the pretraining stage. (c) Reconstructions after the backpropagation. (d) Original image. (e) Fixation density map. (f) Saliency map of the proposed model.

the reconstructions of these patches present obvious dissimilarities with their targets.

According to the relation between reconstruction residual and saliency, to estimate saliency of each pixel, we first infer its local patch by forward propagating the surrounding patch with a larger area through the trained C–S inference network. Then, its saliency can be measured by the reconstruction error of the local patch to pop out the regions with unique C–S pattern which is difficult to obtain and learn. Concretely, for each pixel $\boldsymbol{x} \in \mathbb{R}^2$ in an image, we use $\boldsymbol{s}(\boldsymbol{x}) \in \mathbb{R}^{cD^2}$ and $\boldsymbol{c}(\boldsymbol{x}) \in \mathbb{R}^{cd^2}$ to denote the vectors of the $D \times D$ surrounding and $d \times d$ central patches at $\boldsymbol{x}$, which are formed by stacking all the columns of the patch in $c$ channels and performing the same preprocessing step as training data. With the learned C–S inference network $\boldsymbol{f}(\cdot)$, we denote the reconstructed central patch by the network as $\boldsymbol{f}(\boldsymbol{s}(\boldsymbol{x}))$, and the final saliency of the proposed model $\mathcal{S}(\boldsymbol{x})$ can be computed as

$$\mathcal{S}(\boldsymbol{x}) = \|\boldsymbol{f}(\boldsymbol{s}(\boldsymbol{x})) - \boldsymbol{c}(\boldsymbol{x})\|_2. \quad (5)$$

Although the cross-entropy error is the objective function for optimization in the backpropagation step, experimentally, we observed worse saliency detection results using the cross-entropy error as the evaluating criterion for reconstruction and saliency. The best results were obtained by training with the cross-entropy error while measuring reconstruction with the $\ell_2$-norm of the differences between the predictions and the actual central vectors of pixels.

To conclude, in the proposed model, the region will be regarded as salient if the observation of it differs obviously from the previous prediction made under the prior distribution. Therefore, with residual as the measure, the concept of saliency in this paper can also be interpreted as the unpredictable or surprising content in the images, which is consistent with the definition of surprise from the perspective of information theory [1].

## IV. Experiments

### A. Quantitative Metrics of Assessment

There exist various measures that have widely been used to evaluate saliency models. According to a recent research on the comparison metrics for eye fixation prediction assessment [65], [66], we utilized the following metrics which are

location-based, value-based, and distribution-based, respectively, to derive a fair and comprehensive evaluation of models.

*1) Area Under the Curve:* Area under the curve (AUC) is the area under receiver operating characteristic (ROC) curve. To plot an ROC curve, all fixations in the current image are considered as positive samples with the same number of points extracted uniformly from the nonfixation locations as negative ones. Then, by varying a threshold to binarize the saliency map, the predicted map can be regarded as a binary classifier to separate the positive samples from the negatives, based on which the indicators of the curve can be calculated. For each saliency map, we repeated the above operation 100 times to extract negative samples. The final ROC curve of each model was obtained by averaging the results over 100 random permutations and all the images in the data set.

*2) Shuffled AUC:* As Zhang *et al.* [24] and Tatler [67] have pointed out, human fixations usually have a strong bias toward the center of visual inputs. The so-called center-bias (CB) effect may affect the original AUC evaluation. To achieve a fairer comparison between the models with and without the consideration of CB, we followed the proposal of Zhang *et al.* [24] and Tatler [67], to use the shuffled AUC (sAUC) score, a refined evaluation procedure of the standard AUC. The main difference is that instead of sampling from the nonfixated pixels of the current image, the negative samples are extracted from the union of all the fixations across all the images in the same data set except for the positive points. Due to the property of being robust to CB and border effect, sAUC has been adopted in a lot of research on saliency [18], [24], [65], [66], [68] and becomes one of the key metrics to fairly assess distinct models. Similar to the standard AUC metric, sAUC score will output a scalar value between 0 and 1 with 0.5 for the chance and 1 for the perfect prediction.

*3) Normalized Scanpath Saliency:* Normalized scanpath saliency (NSS) was introduced in [69] to measure the correspondence between estimated saliency values and human fixations. In particular, the saliency map $\mathcal{S}(\boldsymbol{x})$ is first normalized to have zero mean and unit standard deviation, and then along the subjects' scanpath, the NSS score can be computed according to the average of the predicted saliency values at the fixation locations in the normalized map as

$$\mathrm{NSS} = \frac{1}{N}\sum_{i=1}^{N}\frac{\mathcal{S}(\boldsymbol{x}_f^i) - \mu_s}{\sigma_s} \tag{6}$$

where $N$ is the total number of fixation points, and $\mathcal{S}(\boldsymbol{x}_f^i)$ is the saliency score of the $i$th fixated location $\boldsymbol{x}_f^i$. In addition, $\mu_s$ and $\sigma_s$ are the mean and standard deviation of the original saliency map to achieve the normalization, respectively.

*4) Kullback–Leibler Divergence:* KLD is utilized to compute the dissimilarity between two probability density functions and has also been adopted as a metric for saliency assessment [65], [70], [71]. Different from the previous metrics which should be maximized, KLD is utilized as a divergence measure in the evaluation process to provide complementary results. Mathematically, given a saliency map $\mathcal{S}(\boldsymbol{x})$ and its corresponding GT of fixation density map $\mathcal{S}_{\mathrm{GT}}(\boldsymbol{x})$, the values of the comparison metric KLD can be calculated according to

$$\mathrm{KLD} = \sum_{\boldsymbol{x}} \mathcal{S}_{\mathrm{GT}}(\boldsymbol{x}) \log\left(\frac{\mathcal{S}_{\mathrm{GT}}(\boldsymbol{x})}{\mathcal{S}(\boldsymbol{x}) + \epsilon} + \epsilon\right) \tag{7}$$

where $\epsilon$ is a small constant to avoid log and division by zero, and both maps $\mathcal{S}(\boldsymbol{x})$ and $\mathcal{S}_{\mathrm{GT}}(\boldsymbol{x})$ are normalized to the sum of one to form the probability distribution.

Except KLD, which was implemented by us, the source codes for the metrics of AUC, sAUC, and NSS were downloaded from: http://saliency.mit.edu/downloads.html.

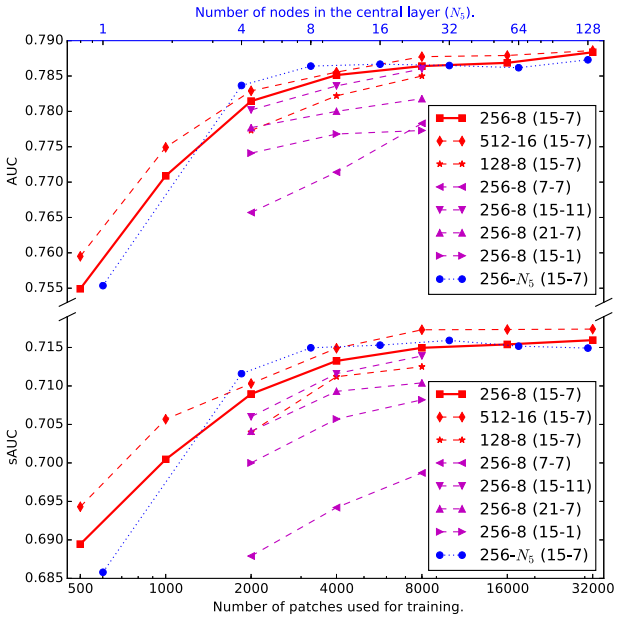


Fig. 4. Effect of parameters on the assessment scores. Red curves: comparison of overall structure. Purple curves: comparison of the size of C–S patches. Blue curves: comparison of the size of central nodes; 256-8 refers to the network that consists of an encoder with the layers of size 256-128-64-32-8, 512-16 refers to a larger network of 512-256-128-64-16, and 128-8 refers to another structure of 128-64-32-16-8 for reference. In addition, 256-$N_5$ denotes the size of encoder as 256-128-64-32-$N_5$ with $N_5$ units in the central layer.

## B. Selection of Parameters

In this section, we investigated the effect of key parameters on the final detection results, which include the number of sampling pixels $m$, the size parameters of the surrounding and central patches $D$–$d$, and the structure of the network. The overall saliency detection performance was measured by the AUC and sAUC scores. A comparison of results with different parameter settings is shown in Fig. 4.

According to the research on deep learning [57], an architecture with enough depth will be easier to train than the ones with fewer layers. For a shallow network, it will require infinite samples to train, while for deep architectures with enough number of layers, they only require finite number of samples. To explore the required number of training samples and to verify the convergence of the model, we first analyzed the relationship between the number of training samples and the metric scores for two networks with different sizes.

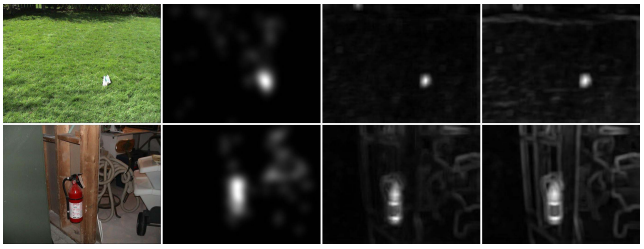

Fig. 5. Comparison between C–S reconstruction and center-only self-reconstruction. From left to right: columns are original images, fixation density maps, and the results of *D–d* being 15–7 and 7–7, respectively.

As can be observed from Fig. 4, the two networks have similar convergence curves. For each network, more sampling data will help the network to capture more information from the current scene and can bring about better performance. Besides, when $m$ is greater than 8000, no much benefit can be gained from the evaluation scores. From the aspect of data quantity, 8000 samples will bring the amount of data of $8000 \times 15 \times 15 \times 3 = 5\,400\,000$, which could provide enough constraints for training (the network of 256-8 is comprised of 533 143 parameters).

With the increment of training samples, the computation time will increase linearly. For instance, for the network structure of 256-8, in our MATLAB implementation, it takes approximately 9 s to process a 511 × 681 image with 500 samples on a computer with a 3.4 GHz Intel i7-2600 CPU, and 17, 34, 69, and 135 s for the cases achieved by doubling training samples gradually.

For the size of C–S patches, in Fig. 4, the results of distinct combinations of *D–d* values were tested to further illustrate the rationality of the selected parameters. *D–d* of 7–7 indicates that the saliency of each region is decided by its self-reconstruction. In this case, although the reconstruction process does not use extra surrounding contrast, the contrast mechanism has also been introduced to the model by global competition, and the trained network has contained the global information. Just because of the sampling process, the repeated background patches can be reconstructed with better results, while the unusual foreground can be popped out. Therefore, the results under the circumstances are still acceptable.

In spite of this, the model of self-reconstruction may present some problems. A comparison between C–S reconstruction and center-only reconstruction is shown in Fig. 5. As can be observed from the figure, the model of self-reconstruction may focus more on the unique patterns, and hence, some edges of the background are also highlighted in the saliency maps. As for C–S inference, these regions may have similar patterns of central and surrounding patches to generate low inference residual and saliency with lower false positive rate. The quantitative results in Fig. 4 can also confirm the significance of contrast to model saliency. According to the comparisons of multiple parameters, in the following experiments, $m$ is set as 8000, and the surrounding and central windows are fixed of 15 × 15 and 7 × 7, respectively, to predict saliency with preferable results and affordable computation time.

To explore the impact of the network structure, in Fig. 4, we first compared the results of two networks with double number of neurons in each hidden layer. As shown in Fig. 4, the larger the network is, the more powerful description it can achieve with better performance, but it needs more samples to train with greater computational cost. In addition, besides the overall structure, we computed the influence of the number of nodes in the central layer for the network structure 256-128-64-32-$N_5$, with the number of central units $N_5$ varying from 1, 4, 8, 16, 32, and 64 to 128. For $N_5$, there is no noticeable change when its size becomes greater than 8. In order to achieve a further analysis, for the cases of 8 and 64 central units, we calculated the average absolute responses of central units in the saliency estimation process on Bruce and Tsotsos' [22] data set (with the results of 3.0812 and 0.8584, respectively), and their average standard deviation on each image (with the results of 3.0385 and 0.7098, respectively). As can be observed from the comparison, although the dimension of features increases, the average response and variability of the units decrease. Hence, the actual ability of the central code layer to represent inputs discriminatingly may not change much with similar results. Generally speaking, for the structure of the network, the larger network with more powerful capacity of description can enhance the upper bound of the detection performance. However, for the central layer, the results may remain stable after its size has reached a certain value. Among these sizes, we finally chose the network structure of 256-128-64-32-8 for the subsequent calculation.

### C. Experiments on Public Data Sets

In this section, we compared the proposed method with the 19 state-of-the-art models, which are denoted as Itti [3], attention based on information maximization (AIM) [22], spectral residual (SR) [43], Gao [13], incremental coding length (ICL) [23], saliency using natural statistics (SUN) [24], Seo [32], image signature (IS) [68], context-aware model (CA) [20], adaptive whitening saliency (AWS) [72], local and global rarity (LG) [18], graph-based visual saliency (GBVS) [73], Judd [31], Wu [16], CovSal [35], Ren [25], Shen [40], nonlocal center-surround reconstruction (NCSR) [15], and NCSR_*g* [17], respectively. The models were selected according to the number of citations [3], [22], [43], performance in the previous comparative studies [18], [23], [31], [68], [72], [73], and relativity to the proposed approach [15]–[17], [25], [40]. The implementations of most of the methods were downloaded from the authors' project Web sites, while the source codes for AWS [72] and Shen [40] were obtained directly from us. The models were used with their default parameters, and the generated saliency maps were resized to the same resolution as the original images in order to compare with the fixation data. For Ren *et al.*'s [25] and Gao and Vasconcelos' [13] methods, we did not have access to their codes; hence, we reimplemented their algorithms with our own MATLAB codes and followed all the parameters they had provided. For instance, in the implementation of [25], we adopted two scales of 80 × 60 and 160 × 120, and set the size of local patches as 8 × 8 with 50% overlap. In addition, the smoothing operation may also affect the scores [68]. In our model, the parameter of Gaussian kernel was set to a simple constant as [68]. Therefore, for the algorithms without the

TABLE I
CHARACTERISTICS OF THE FOUR SELECTED EYE-TRACKING DATA SETS

| Dataset | Number of images | Average viewers | Resolution | Viewing distance(cm) | Viewing time / Interval(s) | Remarks |
|---|---|---|---|---|---|---|
| DS1 [22] | 120 | 20 | 681×511 | 75 | 4/2 | The most widely used dataset including outdoor and indoor scenes |
| DS2 [31] | 1003 | 15 | Various | 60 | 3/1 | The largest dataset with 779 landscape images and 228 portrait images |
| DS3 [33] | 101 | 31 | 1024×768 | 70 | N/A | One of the datasets with the highest number of eye-tracking subjects |
| DS4 [59] | 758 | 25 | Various | 76 | 5/2 | The dataset containing a large number of semantically affective images |

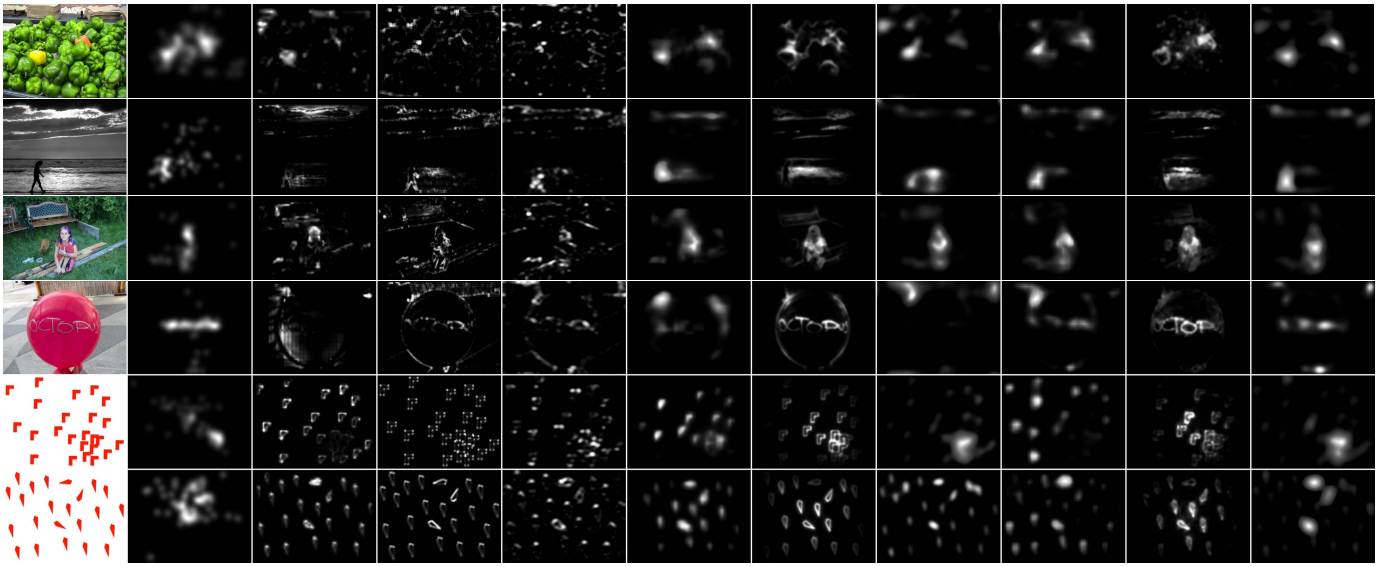


Fig. 6. Visual comparison of histogram matched saliency maps with the GT. Top one, middle one, and bottom four rows are images from DS1 [22], DS4 [59], and DS2 [31], respectively.

smoothing parameters, we offered the best evaluation results of their models on a set of smoothing parameters, to achieve a fairer comparison with others.

Four widely used public data sets, which are provided by Bruce and Tsotsos [22], Judd *et al*. [31], Kootstra *et al*. [33], and Ramanathan *et al*. [59], respectively, were selected in the benchmark because of their distinct stimulus categories and eye-tracking parameters, with the details listed in Table I.

For a qualitative evaluation, we presented the visual comparison of different saliency maps on the benchmark data sets. Directly generated saliency maps of distinct methods are markedly different from each other in the amount of salient regions. Therefore, to equalize the amount of salient pixels returned by different approaches, the method in [38] and [74] is utilized to histogram match the saliency maps to the corresponding fixation density maps. With better highlights on most salient locations, the results of saliency maps are shown in Fig. 6.

Take a typical visual search task in the last row, for example. For the nonlocal-based methods [16], [17], which show inspiring results in the most cases, they generate similar results for the target and distracting items in the nonlocal comparison due to practically limited nonlocal regions. Therefore, they fail to pop out the target items. However, by training the C–S relationship in a global manner, our method emphasizes more on the learning and modeling of the distracting items. Thus, with less learning and more information, the target items win the competition of calculating reconstruction residual to the present saliency.

To compare the proposed method with others quantitatively, the standard ROC curves of the saliency models on the four data sets are shown in Fig. 7, with the corresponding AUC scores listed in Table II. Besides, the results under other three metrics are also listed in Table II to provide a fair comparison. It is noteworthy that to achieve a clearer presentation, in Table II, the models were classified into two groups according to whether they had integrated a CB item or not. In addition, we provided the results of the proposed method with and without the CB item for comparison. To generate the center-weighted version of the proposed model (AER + CB), the CB model in [17] was used. As can be observed from the results, the standard AUC and NSS may expect the algorithms to model the CB, while sAUC penalizes the models with CB items in that these methods will generate higher false positive rate under sAUC metric. Regarding KLD metric, as an outlier different from most of the other metrics [65], neither of the groups has obvious advantages.

With the evaluation results of diverse metrics on distinct data sets, to conduct a more comprehensive evaluation, Borda count (BC) election method was used. BC was chosen

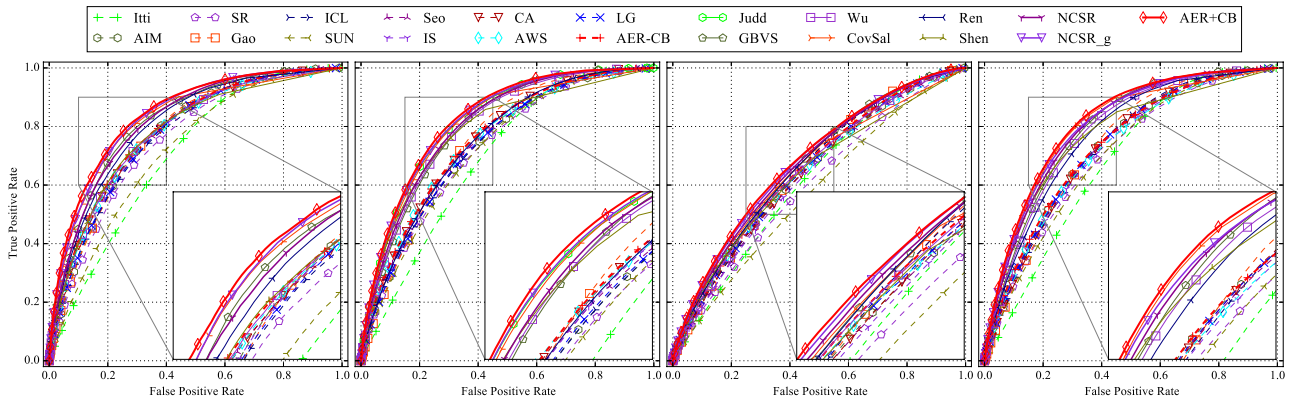


Fig. 7. Standard ROC curves on the four benchmark data sets. From left to right: data sets are DS1 [22], DS2 [31], DS3 [33], and DS4 [59], respectively. Solid lines: models with CB. Dotted lines: models without CB. To derive the ROC curves, the human fixations of the current image were regarded as the positive samples, and the same number of points was sampled uniformly from the nonfixation locations to form the negative set.

TABLE II

COMPARISON OF QUANTITATIVE SCORES OF DIFFERENT SALIENCY MODELS. TOP GROUP: MODELS WITHOUT CB ITEMS. BOTTOM GROUP: MODELS MODELING THE CB IN THEIR SALIENCY MAPS. AER – CB REFERS TO THE PROPOSED AUTOENCODER RECONSTRUCTION-BASED MODEL WITHOUT CB, AND AER + CB REFERS TO A CENTER-WEIGHTED VERSION OF THE PROPOSED METHOD

| Algorithm | DS1 [22] | | | | DS2 [31] | | | | DS3 [33] | | | | DS4 [59] | | | | BC |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AUC | sAUC | NSS | KLD | AUC | sAUC | NSS | KLD | AUC | sAUC | NSS | KLD | AUC | sAUC | NSS | KLD | |
| Itti [3] | 0.702 | 0.567 | 0.742 | 1.272 | 0.692 | 0.568 | 0.687 | 1.721 | 0.620 | 0.546 | 0.450 | 0.702 | 0.681 | 0.526 | 0.638 | 1.407 | 25 |
| AIM [22] | 0.775 | 0.694 | 1.086 | 1.137 | 0.732 | 0.682 | 1.011 | 1.594 | 0.640 | 0.594 | 0.530 | 0.689 | 0.732 | 0.645 | 0.955 | 1.297 | 108 |
| SR [43] | 0.751 | 0.673 | 0.964 | 1.205 | 0.723 | 0.653 | 0.819 | 1.693 | 0.601 | 0.581 | 0.399 | 0.866 | 0.706 | 0.626 | 0.821 | 1.386 | 45 |
| Gao [13] | 0.773 | 0.657 | 1.219 | 1.011 | 0.752 | 0.664 | 1.049 | **1.454** | 0.653 | 0.558 | 0.571 | 0.659 | 0.734 | 0.615 | 0.918 | 1.299 | 114 |
| ICL [23] | 0.770 | 0.683 | 1.281 | 1.046 | 0.735 | 0.671 | 1.077 | 1.471 | 0.653 | 0.593 | 0.601 | 0.699 | 0.729 | 0.617 | 0.926 | 1.310 | 111 |
| SUN [24] | 0.717 | 0.662 | 0.915 | 1.219 | 0.722 | 0.665 | 0.913 | 1.635 | 0.576 | 0.559 | 0.338 | 0.781 | 0.703 | 0.618 | 0.805 | 1.354 | 39 |
| Seo [32] | 0.782 | 0.704 | 1.200 | 1.035 | 0.743 | 0.674 | 0.947 | 1.567 | 0.633 | 0.597 | 0.497 | 0.710 | 0.728 | 0.625 | 0.864 | 1.298 | 98 |
| IS [68] | 0.780 | 0.698 | **1.379** | **1.008** | 0.748 | 0.677 | 1.042 | 1.532 | 0.633 | 0.598 | 0.559 | 0.665 | 0.722 | 0.633 | 0.943 | **1.257** | 132 |
| CA [20] | 0.780 | 0.695 | 1.307 | 1.034 | 0.753 | 0.688 | 1.071 | 1.519 | 0.642 | 0.601 | 0.573 | 0.705 | 0.730 | 0.628 | 0.903 | 1.298 | 132 |
| AWS [72] | 0.778 | 0.712 | 1.263 | 1.031 | 0.751 | 0.699 | 1.100 | 1.498 | 0.645 | **0.620** | 0.599 | 0.740 | 0.725 | **0.649** | 0.950 | 1.303 | 143 |
| LG [18] | 0.773 | 0.693 | 1.142 | 1.107 | 0.735 | 0.691 | 1.023 | 1.563 | 0.649 | 0.598 | 0.583 | 0.670 | 0.731 | 0.639 | 0.933 | 1.286 | 120 |
| AER-CB | **0.786** | **0.715** | 1.322 | 1.034 | **0.754** | **0.702** | **1.104** | 1.512 | **0.655** | 0.614 | **0.644** | **0.653** | **0.735** | 0.646 | **0.974** | 1.261 | 181 |
| GBVS [73] | 0.819 | 0.641 | 1.519 | 0.852 | 0.799 | 0.679 | 1.308 | 1.297 | 0.667 | 0.559 | 0.632 | 0.684 | 0.793 | 0.626 | 1.233 | 1.036 | 72 |
| Judd [31] | N/A | N/A | N/A | N/A | 0.818 | 0.683 | 1.283 | 1.547 | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A | N/A |
| Wu [16] | 0.782 | 0.643 | 1.442 | 1.674 | 0.798 | 0.673 | 1.277 | 1.359 | 0.650 | 0.565 | 0.577 | 0.871 | 0.783 | 0.608 | 1.145 | 1.130 | 44 |
| CovSal [35] | 0.821 | 0.592 | 1.495 | 1.230 | 0.805 | 0.638 | 1.403 | 1.729 | 0.655 | 0.531 | 0.672 | 3.475 | 0.799 | 0.584 | 1.279 | 1.912 | 50 |
| Ren [25] | 0.805 | 0.669 | 1.504 | 1.045 | 0.748 | 0.679 | 1.318 | 1.305 | 0.665 | 0.579 | 0.642 | 0.910 | 0.771 | 0.617 | 1.096 | 1.376 | 61 |
| Shen [40] | 0.777 | 0.620 | 1.303 | 1.521 | 0.784 | 0.641 | 1.389 | 1.929 | 0.637 | 0.537 | 0.605 | 2.116 | 0.770 | 0.601 | 1.255 | 1.611 | 33 |
| NCSR [15] | 0.818 | 0.650 | 1.525 | 0.892 | 0.806 | 0.679 | 1.330 | 1.373 | 0.673 | 0.561 | 0.679 | 0.621 | 0.800 | 0.616 | 1.244 | 1.103 | 79 |
| NCSR_$g$ [17] | 0.837 | 0.676 | 1.801 | 0.715 | 0.821 | 0.683 | 1.511 | 1.179 | 0.680 | 0.568 | 0.747 | 0.663 | 0.803 | 0.618 | 1.339 | 1.028 | 109 |
| AER+CB | **0.842** | **0.690** | **1.805** | **0.670** | **0.826** | **0.684** | **1.538** | **1.111** | **0.684** | **0.584** | **0.771** | **0.586** | **0.815** | **0.634** | **1.388** | **0.941** | 128 |

here for the reason that it can elect the broadly acceptable options and is described as the consensus-based voting system. In other words, under such circumstances, a good model should have relatively high scores on most of the metrics if it is not the best. The BC scores of the two groups were calculated separately. Concretely, for each group, we took each column in Table II as a voter and assigned points for the methods according to their rankings. For instance, for the comparison among models without CB, under the simplest form of BC, the method being ranked first received 12 points, the second 11 points, and so on, with the last one point. Then, for each method, we added up the points received from different metrics and data sets to generate the final points, which are listed in the last column in Table II. As shown in the comparison, for the group without CB, the proposed method (AER–CB) can win the competition and yield comparable results to the state-of-the-art models in the comparative studies, such as AWS [72] and LG [18]. For the other group, our reconstruction-based models and GBVS [73] can present a better performance.

Built on the finding that the HVS is well adapted for the efficient coding of information in the visual environment, many computational saliency models [18], [22]–[24], [40] have been proposed. As a model deriving features for saliency estimation with a learned encoding process of patches, a comparison with other algorithms related to efficient encoding is required. In the previous models [18], [22]–[24], a suitable basis is usually obtained by learning from a group of images with natural scenes, and features are extracted based on the basis functions. In spite of similar features, as shown in Table II, the models with different contexts to highlight sparsity show distinct performance. For instance, the algorithms computing the probability distribution and estimating the rarity from the current image [22], [23] can outperform the one gaining the

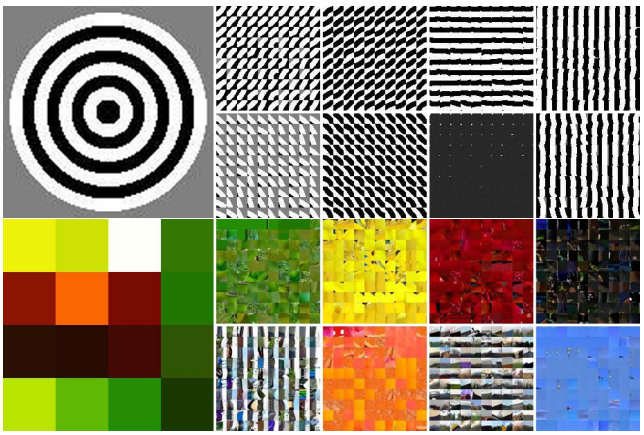

Fig. 8. Visualization results. Given an original image (left), we can derive the most responsive stimuli of eight central units, respectively (right). Top: we calculated the central layer responses of all the patches from the input image itself and presented 100 patches with the highest responsive values of the eight hidden units, respectively. Bottom: image patches (about 300 000) extracted from the other data sets were utilized for comparison.

statistics of features from another set of natural images [24], which indicate the underlying importance of exploring the effective information from the current scenes for saliency modeling. In addition, the model combining local and global consideration, such as the proposed method and LG [18], can exhibit a better performance than the models with an emphasis merely on global rarity [22], [23].

Distinct from the existing approaches which generate features with fixed basis [18], [22]–[24], the proposed model can derive the encoding features adaptively for distinct visual inputs. Furthermore, after the training process, our model can also obtain an optimal description of the C–S contrast under each scene instead of using human-designed C–S contrast calculation mechanisms. Compared with LG [18] which measures local saliency by the average-weighted feature dissimilarities between a center patch and others in the neighborhood, the proposed method with the C–S inference network modulated by the current image content can achieve a better detection of salient regions for different types of images.

### *D. Visualizing Learned Features*

In this section, a simple visualization technique [54] was utilized to illustrate what the deep network has learned after training. In other words, we tested whether the central code layer of the network can learn certain representative and meaningful features of the high-dimensional inputs with a limited number of hidden units.

In particular, given the C–S reconstruction network trained by the current image, we computed the activations of the central units on a large set of patches with various patterns. Then, for each of the eight units, we sorted all the patches according to their activation values of the selected unit and displayed the top-scoring patches to find out what kind of patches it fires on under different conditions. As shown in Fig. 8, at the top, we calculated the central layer responses of all the patches from the input image itself and presented 100 patches with the highest responsive values of the eight hidden units, respectively, while at the bottom, image patches (about 300 000) extracted from other data sets [8], [22], [31] were utilized for comparison to further test the ability of the deep C–S inference network to learn features.

As can be seen from the visualization results, unlike the fixation-based saliency prediction models [39], [40] which emphasize more on the high-level semantic features like faces and texts, the proposed network appears to learn some significant low-level features in the HVS [2] for saliency modeling, such as local orientation, color, and brightness. Furthermore, instead of using the identical features in all the cases [3], [13], the deep network trained with the input image can discover visual features suitable for describing the C–S relationship under current visual stimulus adaptively and, therefore, can achieve a better estimation of C–S contrast and saliency.

## V. Discussion

In the proposed model, the autoencoder network can simulate the layer-by-layer abstraction and propagation of information in the human visual cortex [27]–[29]. In addition, the C–S reconstruction can simulate the prediction process in the sensory cortex to stress the unexpected data [1], [42]. Based on the deep C–S inference architecture, the HVS can be considered as a control system with the total reconstruction residual as the feedback signal to control eye movements. In the early stage of visual processing, eye movements are distributed on the entire visual field to obtain a rough representation of scenes, which is akin to the initial random saccades in viewing a new image in eye-tracking experiments. With scene information captured at a coarse scale, the residual of each region can be estimated to provide an initial residual map of surprise degree and the total residual of the scene.

According to the learning process of autoencoder networks, regions sampled with high weights tend to present small reconstruction residual. Therefore, in the second stage, to decrease the total reconstruction residual of the scene, the system will shift fixations to the most salient regions with the largest residual sequentially, in a way of biologically-plausible winner-take-all (WTA) and inhibition of return [3]. As a result, the HVS will tend to quiet down and reach the steady state of the lowest residual and uncertainty. With this assumption of residual feedback modulation mechanism, a plausible explanation of visual saliency can be proposed to establish a bridge between the random saccades stage and the subsequent shifts of fixations in a WTA manner [3], which we hope can provide more inspiration for the further research on bottom–up saliency modeling. The assumption is also consistent with the predictive coding theory [75], [76], which is believed to be the design principle of the nervous system. According to this theory, the reconstruction error, termed prediction error, controls the information reduction and transmission in the nervous system.

Over the past decade, selecting and weighting features have always been important issues for saliency estimation. The early studies may focus more on exploring the visual features that guide the deployment of saliency [2], [3], and in the research of recent years, determining how to weight the contribution of features has also attracted increasing interest

TABLE III
LSUN CHALLENGE RESULTS. OUR RESULTS, UNDER THE TEAM NAME XIDIAN, WERE OBTAINED BY APPLYING THE AER − CB MODEL. THE TOP TWO RESULTS ARE HIGHLIGHTED IN BOLD FONT

| Algorithm | iSUN | | SALICON | |
|---|---|---|---|---|
| | AUC | sAUC | AUC | sAUC |
| Itti [3] | 0.7262 | 0.6024 | 0.6603 | 0.6101 |
| GBVS [73] | 0.7913 | 0.6208 | 0.7816 | 0.6303 |
| BMS [77] | 0.6560 | 0.5885 | 0.7699 | **0.6935** |
| UPC [78] | **0.8463** | **0.6650** | **0.8291** | 0.6698 |
| **Xidian** (AER-CB) | 0.7949 | **0.6484** | 0.7990 | **0.6809** |
| WHU_IIP | **0.7960** | 0.6307 | 0.7759 | 0.6064 |
| LCYLab | 0.7921 | 0.6259 | N/A | N/A |
| Rare Improved | 0.7582 | 0.6283 | **0.8047** | 0.6644 |

with the expansion of feature dimension. In particular, to fit the sophisticated relationship between various features and saliency, learning-based algorithms [31], [34], [40], [47] have emerged by learning high-level information from known fixations and shown a better performance than the traditional models based on low-level cues. More recently, CNNs have been applied with promising results to saliency detection. In [79], a CNN-based saliency model has been presented to learn features from roughly half of Judd *et al.*'s [31] data set, which can significantly outperform the previous methods on that data set. Another example can be found from Table III. In the large-scale scene understanding (LSUN) challenge,[1] the CNN-based model proposed by the team UPC [78] illustrates the best performance in the saliency prediction challenge.

Although the recent success of supervised learning has overshadowed unsupervised learning, the research on learning from the unlabeled data has also shown its significance in understanding human learning. As shown in Table III, under sAUC, the relatively fair metric, our unsupervised model (Xidian) can produce results competitive to the supervised method, UPC [78]. Note that our results were obtained by training the reconstruction network with only the input image and not using the annotated data provided in the LSUN, while UPC's results rely on 6000 and 10 000 annotated image pairs to train the model of respective data set. In addition, in Table II, as a data-driven model, our method can exhibit a better capability to predict fixations than the fixation-based algorithms [31], [40], even on the data sets with more emphasis on semantic objects [31], [59]. Thus, instead of an opposite concept, stimulus-driven bottom–up saliency may also be important for the research on task-dependent attention [30] and other high-level applications, such as object-based saliency detection and segmentation [80].

In saliency estimation, C–S contrast is one of the most important ingredients [3], [4]. To measure C–S contrast, models have evolved from pixel-level information processing [3], [12] to patch-based C–S comparison [14]–[16], [25]. Based on the observation that the same C–S region may show distinct saliency under different environments [30], image-level global information has been introduced to measure C–S contrast [17]–[20]. With global uniqueness, the estimation quality is greatly improved, which further verifies the significance of global integration in C–S contrast computation. From a converse aspect, in the comparison of global algorithms, the ones based on C–S contrast [19], [20] can integrate the local spatial relationships into the global competition. Therefore, as indicated in Table II, C–S comparison can present more efficient ability to measure global rarity than other schemes [23], [24], [43]. In addition, psychophysical experiments [81] on V1 saliency also show the significance of combining local error detection based on C–S kernels with global comparison.

Behavioral studies [82] show that visual input is processed in a coarse-to-fine fashion. The fast global perception is conveyed in the brain to guide or influence the subsequent more detailed and slower local processing. Therefore, in contrast to approaches incorporating global rarity [17], [18] in postprocessing, the proposed model, which incorporates global competition into the nonlocal computation of low-level areas, is in more agreement with the human visual perception [82].

Compared with human fixations, the proposed model has some limitations. First, for large smooth objects, as many existing bottom–up saliency methods, the proposed AER model may fail to generate satisfactory results and tend to highlight the regions near the object's boundary. This is principally because the hypothesis that background regions present a higher probability of being sampled is not valid in these circumstances. Therefore, modeling the background to reinforce the information acquisition from the background regions can be one of the directions for future research. Second, while the proposed model has presented a capability to detect saliency on the data sets with high-level objects, as mentioned earlier, studying how to integrate high-level knowledge to pop out the semantic regions more consistent with human fixations, such as faces, pedestrian, and cars, is also a problem worthy of further investigation in the future.

## VI. CONCLUSION

In this paper, we have proposed a deep autoencoder-based C–S inference network to model the human visual perception process and to estimate bottom–up saliency. With global competition in sampling and learning, different regions present distinct reconstruction ability and C–S relationship under the unified reconstruction parameters, according to which saliency is assessed. By integrating global rarity into the computation of local C–S contrast, the proposed method can perform better than the models with individual consideration and show competitive results to the state-of-the-art algorithms. In addition, distinct from the previous combination ways of local and global information, it can explore a novel view to address the saliency detection problem in the research direction encouraged by [18].

## ACKNOWLEDGMENT

The authors would like to thank Y. Huang and C. Shen for their assistance in conducting the experiments, and C. Shen and X. R. Fdez-Vidal for sharing their codes. They would also like to thank the editors and all the anonymous reviewers for their helpful comments and valuable suggestions on the previous versions of this paper.

[1]LSUN leaderboard: http://lsun.cs.princeton.edu/leaderboard/

## REFERENCES


[1] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vis. Res.*, vol. 49, no. 10, pp. 1295–1306, Jun. 2009.

[2] J. M. Wolfe and T. S. Horowitz, "What attributes guide the deployment of visual attention and how do they do it?" *Nature Rev. Neurosci.*, vol. 5, no. 6, pp. 495–501, 2004.

[3] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.

[4] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, Mar. 2001.

[5] Z. Li, "A saliency map in primary visual cortex," *Trends Cognit. Sci.*, vol. 6, no. 1, pp. 9–16, Jan. 2002.

[6] L. Itti, "Automatic foveation for video compression using a neurobiological model of visual attention," *IEEE Trans. Image Process.*, vol. 13, no. 10, pp. 1304–1318, Oct. 2004.

[7] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Trans. Image Process.*, vol. 19, no. 1, pp. 185–198, Jan. 2010.

[8] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.

[9] D. Erhan, C. Szegedy, A. Toshev, and D. Anguelov, "Scalable object detection using deep neural networks," in *Proc. 27th IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 2155–2162.

[10] D. Gao and N. Vasconcelos, "Discriminant saliency for visual recognition from cluttered scenes," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2004, pp. 481–488.

[11] A. A. Salah, E. Alpaydin, and L. Akarun, "A selective attention-based method for visual pattern recognition with application to handwritten digit recognition and face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 420–425, Mar. 2002.

[12] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proc. 11th ACM Int. Conf. Multimedia*, Berkeley, CA, USA, Nov. 2003, pp. 374–381.

[13] D. Gao and N. Vasconcelos, "Bottom-up saliency is a discriminant process," in *Proc. 11th IEEE Int. Conf. Comput. Vis.*, Rio de Janeiro, Brazil, Oct. 2007, pp. 185–190.

[14] D. A. Klein and S. Frintrop, "Center-surround divergence of feature statistics for salient object detection," in *Proc. 13th IEEE Int. Conf. Comput. Vis.*, Barcelona, Spain, Nov. 2011, pp. 2214–2219.

[15] C. Xia, P. Wang, F. Qi, and G. Shi, "Nonlocal center-surround reconstruction-based bottom-up saliency estimation," in *Proc. 20th IEEE Int. Conf. Image Process.*, Melbourne, VIC, Australia, Sep. 2013, pp. 206–210.

[16] J. Wu, F. Qi, G. Shi, and Y. Lu, "Non-local spatial redundancy reduction for bottom-up saliency estimation," *J. Vis. Commun. Image Represent.*, vol. 23, no. 7, pp. 1158–1166, Oct. 2012.

[17] C. Xia, F. Qi, G. Shi, and P. Wang, "Nonlocal center–surround reconstruction-based bottom-up saliency estimation," *Pattern Recognit.*, vol. 48, no. 4, pp. 1337–1348, Apr. 2015.

[18] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 478–485.

[19] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 409–416.

[20] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[21] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu, "Visual saliency detection by spatially weighted dissimilarity," in *Proc. 24th IEEE Conf. Comput. Vis. Pattern Recognit.*, Colorado Springs, CO, USA, Jun. 2011, pp. 473–480.

[22] N. D. B. Bruce and J. K. Tsotsos, "Saliency based on information maximization," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, Dec. 2005, pp. 155–162.

[23] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Proc. Annu. Conf. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2008, pp. 681–688.

[24] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, pp. 1–20, Dec. 2008.

[25] Z. Ren, S. Gao, L.-T. Chia, and D. Rajan, "Regularized feature reconstruction for spatio-temporal saliency detection," *IEEE Trans. Image Process.*, vol. 22, no. 8, pp. 3120–3132, Aug. 2013.

[26] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proc. 14th IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 2976–2983.

[27] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[28] Q. V. Le *et al.*, "Building high-level features using large scale unsupervised learning," in *Proc. 29th Int. Conf. Mach. Learn.*, Edinburgh, Scotland, Jun. 2012, pp. 1–11.

[29] A. Krizhevsky and G. E. Hinton, "Using very deep autoencoders for content-based image retrieval," in *Proc. 19th Eur. Symp. Artif. Neural Netw.*, Bruges, Belgium, Apr. 2011, pp. 1–7.

[30] A. Torralba, A. Oliva, M. S. Castelhano, and J. M. Henderson, "Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search," *Psychol. Rev.*, vol. 113, no. 4, pp. 766–786, 2006.

[31] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. 12th IEEE Int. Conf. Comput. Vis.*, Kyoto, Japan, Sep. 2009, pp. 2106–2113.

[32] H. J. Seo and P. Milanfar, "Static and space-time visual saliency detection by self-resemblance," *J. Vis.*, vol. 9, no. 12, pp. 1–27, 2009.

[33] G. Kootstra, A. Nederveen, and B. de Boer, "Paying attention to symmetry," in *Proc. 19th Brit. Mach. Vis. Conf.*, Leeds, U.K., Sep. 2008, pp. 1115–1125.

[34] A. Borji, "Boosting bottom-up and top-down visual features for saliency estimation," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Providence, RI, USA, Jun. 2012, pp. 438–445.

[35] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *J. Vis.*, vol. 13, no. 4, pp. 1–20, Mar. 2013.

[36] Q. Zhao and C. Koch, "Learning saliency-based visual attention: A review," *Signal Process.*, vol. 93, no. 6, pp. 1401–1407, Jun. 2013.

[37] Y. Lu, W. Zhang, C. Jin, and X. Xue, "Learning attention map from images," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 1067–1074.

[38] E. Vig, M. Dorr, and D. Cox, "Large-scale optimization of hierarchical features for saliency prediction in natural images," in *Proc. 27th IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 2798–2805.

[39] C. Shen, M. Song, and Q. Zhao, "Learning high-level concepts by training a deep network on eye fixations," in *Proc. NIPS Deep Learn. Unsupervised Feature Learn. Workshop*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1–8.

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

[41] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 1, pp. 185–207, Jan. 2013.

[42] B. A. Olshausen and D. J. Field, "Emergence of simple-cell receptive field properties by learning a sparse code for natural images," *Nature*, vol. 381, no. 6583, pp. 607–609, 1996.

[43] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. 20th IEEE Conf. Comput. Vis. Pattern Recognit.*, Minneapolis, MN, USA, Jun. 2007, pp. 1–8.

[44] P. Siva, C. Russell, T. Xiang, and L. Agapito, "Looking beyond the image: Unsupervised learning for object saliency and detection," in *Proc. 26th IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 3238–3245.

[45] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, Jan. 1980.

[46] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. 25th IEEE Conf. Comput. Vis. Pattern Recognit.*, Providence, RI, USA, Jun. 2012, pp. 853–860.

[47] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. 26th IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 2083–2090.

[48] Q. Zhao and C. Koch, "Learning a saliency map using fixated locations in natural scenes," *J. Vis.*, vol. 11, no. 3, pp. 1–15, 2011.

[49] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Annu. Conf. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2012, pp. 1106–1114.
[50] P. Sermanet, K. Kavukcuoglu, S. Chintala, and Y. LeCun, "Pedestrian detection with unsupervised multi-stage feature learning," in *Proc. 26th IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 3626–3633.
[51] M. Osadchy, Y. Le Cun, and M. L. Miller, "Synergistic face detection and pose estimation with energy-based models," *J. Mach. Learn. Res.*, vol. 8, pp. 1197–1215, Jan. 2007.
[52] D. Ciresan, U. Meier, J. Masci, and J. Schmidhuber, "A committee of neural networks for traffic sign classification," in *Proc. Int. Joint Conf. Neural Netw.*, San Jose, CA, USA, Jul. 2011, pp. 1918–1921.
[53] C. Szegedy, A. Toshev, and D. Erhan, "Deep neural networks for object detection," in *Proc. Adv. Neural Inf. Process. Syst.*, Lake Tahoe, NV, USA, Dec. 2013, pp. 2553–2561.
[54] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proc. 27th IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 580–587.
[55] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *Int. J. Robot. Res.*, vol. 34, nos. 4–5, pp. 705–724, 2015.
[56] V. Mnih and G. E. Hinton, "Learning to detect roads in high-resolution aerial images," in *Proc. 11th Eur. Conf. Comput. Vis.*, Heraklion, Greece, Sep. 2010, pp. 210–223.
[57] Y. Bengio, "Learning deep architectures for AI," *Found. Trends Mach. Learn.*, vol. 2, no. 1, pp. 1–127, 2009.
[58] B. A. Olshausen and D. J. Field, "Sparse coding with an overcomplete basis set: A strategy employed by V1?" *Vis. Res.*, vol. 37, no. 23, pp. 3311–3325, Dec. 1997.
[59] S. Ramanathan, H. Katti, N. Sebe, M. Kankanhalli, and T.-S. Chua, "An eye fixation database for saliency detection in images," in *Proc. 11th Eur. Conf. Comput. Vis.*, Heraklion, Greece, Sep. 2010, pp. 30–43.
[60] M. D. Zeiler, G. W. Taylor, and R. Fergus, "Adaptive deconvolutional networks for mid and high level feature learning," in *Proc. 13th IEEE Int. Conf. Comput. Vis. (ICCV)*, Barcelona, Spain, Nov. 2011, pp. 2018–2025.
[61] J. van de Weijer, T. Gevers, and J.-M. Geusebroek, "Edge and corner detection by photometric quasi-invariants," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 4, pp. 625–630, Apr. 2005.
[62] G. E. Hinton, "Learning multiple layers of representation," *Trends Cognit. Sci.*, vol. 11, no. 10, pp. 428–434, Oct. 2007.
[63] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.
[64] G. Hinton, "A practical guide to training restricted Boltzmann machines," Dept. Comput. Sci., Univ. Toronto, Toronto, ON, Canada, Tech. Rep. UTML-TR-2010-003, Aug. 2010.
[65] 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 *Proc. 14th IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 1153–1160.
[66] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *IEEE Trans. Image Process.*, vol. 22, no. 1, pp. 55–69, Jan. 2013.
[67] B. W. Tatler, "The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions," *J. Vis.*, vol. 7, no. 14, pp. 1–17, 2007.
[68] X. Hou, J. Harel, and C. Koch, "Image signature: Highlighting sparse salient regions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 1, pp. 194–201, Jan. 2012.
[69] R. J. Peters, A. Iyer, L. Itti, and C. Koch, "Components of bottom-up gaze allocation in natural images," *Vis. Res.*, vol. 45, no. 8, pp. 2397–2416, 2005.
[70] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vis. Res.*, vol. 45, no. 5, pp. 643–659, Mar. 2005.
[71] J. Wang, M. P. DaSilva, P. LeCallet, and V. Ricordel, "Computational model of stereoscopic 3D visual saliency," *IEEE Trans. Image Process.*, vol. 22, no. 6, pp. 2151–2165, Jun. 2013.
[72] A. Garcia-Diaz, V. Leborán, X. R. Fdez-Vidal, and X. M. Pardo, "On the relationship between optical variability, visual saliency, and eye fixations: A computational approach," *J. Vis.*, vol. 12, no. 6, pp. 1–22, 2012.
[73] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proc. Adv. Neural Inf. Process. Syst.*, Vancouver, BC, Canada, Dec. 2006, pp. 545–552.
[74] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," Comput. Sci. Artif. Intell. Lab., Massachusetts Inst. Technol., Cambridge, MA, USA, Tech. Rep. MIT-CSAIL-TR-2012-001, Jan. 2012.
[75] R. P. N. Rao and D. H. Ballard, "Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive-field effects," *Nature Neurosci.*, vol. 2, no. 1, pp. 79–87, 1999.
[76] Y. Huang and R. P. N. Rao, "Predictive coding," *Wiley Interdiscipl. Rev., Cognit. Sci.*, vol. 2, no. 5, pp. 580–593, Sep./Oct. 2011.
[77] J. Zhang and S. Sclaroff, "Saliency detection: A Boolean map approach," in *Proc. 14th IEEE Int. Conf. Comput. Vis.*, Sydney, NSW, Australia, Dec. 2013, pp. 153–160.
[78] J. Pan and X. Giró-i-Nieto. (Jul. 2015). "End-to-end convolutional network for saliency prediction." [Online]. Available: http://arxiv.org/abs/1507.01422
[79] M. Kümmerer, L. Theis, and M. Bethge. (Apr. 2015). "Deep gaze I: Boosting saliency prediction with feature maps trained on ImageNet." [Online]. Available: http://arxiv.org/abs/1411.1045
[80] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. 27th IEEE Conf. Comput. Vis. Pattern Recognit.*, Columbus, OH, USA, Jun. 2014, pp. 280–287.
[81] M. W. Spratling, "Predictive coding as a model of the V1 saliency map hypothesis," *Neural Netw.*, vol. 26, pp. 7–28, Feb. 2012.
[82] V. Beaucousin, G. Simon, M. Cassotti, A. Pineau, O. Houdé, and N. Poirel, "Global interference during early visual processing: ERP evidence from a rapid global/local selective task," *Front. Psychol.*, vol. 4, pp. 1–6, 2013, Art. ID 539.

**Chen Xia** received the B.Eng. degree in electronic information engineering from Xidian University, Xi'an, China, in 2010, where she is currently pursuing the Ph.D. degree in intelligence information processing.

Her current research interests include computer vision and modeling of visual saliency.

**Fei Qi** (M'08) received the B.Eng. degree from Northwestern Polytechnical University, Xi'an, China, in 2000, and the M.Eng. and Ph.D. degrees from Tsinghua University, Beijing, China, in 2007.

He is currently an Associate Professor with the School of Electronic Engineering, Xidian University, Xi'an. His current research interests include bioinspired image/video processing, structural data analysis, and applications of convex optimization.

**Guangming Shi** (SM'10) received the B.S. degree in automatic control, the M.S. degree in computer control, and the Ph.D. degree in electronic information technology from Xidian University, Xi'an, China, in 1985, 1988, and 2002, respectively.

He has been a Professor with the School of Electronic Engineering, Xidian University, since 2003, where he was the Head of National Instruction Base of Electrician and Electronic in 2004. His current research interests include compressed sensing, theory and design of multirate filter banks, image denoising, low-bit-rate image/video coding, and implementation of algorithms for intelligent signal processing.