Toward Statistical Modeling of Saccadic Eye-Movement and Visual Saliency

Xiaoshuai Sun, Hongxun Yao, *Member IEEE*, Rongrong Ji, *Senior Member IEEE*, and Xian-Ming Liu

***Abstract*— In this paper, we present a unified statistical framework for modeling both saccadic eye movements and visual saliency. By analyzing the statistical properties of human eye fixations on natural images, we found that human attention is sparsely distributed and usually deployed to locations with abundant structural information. This observations inspired us to model saccadic behavior and visual saliency based on super-Gaussian component (SGC) analysis. Our model sequentially obtains SGC using projection pursuit, and generates eye movements by selecting the location with maximum SGC response. Besides human saccadic behavior simulation, we also demonstrated our superior effectiveness and robustness over state-of-the-arts by carrying out dense experiments on synthetic patterns and human eye fixation benchmarks. Multiple key issues in saliency modeling research, such as individual differences, the effects of scale and blur, are explored in this paper. Based on extensive qualitative and quantitative experimental results, we show promising potentials of statistical approaches for human behavior research.**

***Index Terms*— Saccadic eye-movement, visual attention, saliency, super Gaussian component analysis.**

I. Introduction

ATTENTION guided saccadic eye-moment is one of the most important mechanisms in biological vision system, based on which the viewer is able to actively explore the environment with high resolution fovea sensors. Benefiting from such unique behavior, human beings, as well as most primates, are able to efficiently process the information from complex environments. For the last four decades, extensive research works have been done by means of theoretical reasoning and computational modeling, trying to uncover the principles that underlie the deployment of gaze. Compared with theoretic hypotheses, computational models of visual attention and saccadic eye-movement not only help us better understand the mechanism of human cognitive behavior but also provide us powerful tools to solve various vision related problems such as video compression [1], scene understanding [2], object detection and recognition [3] *etc.*.

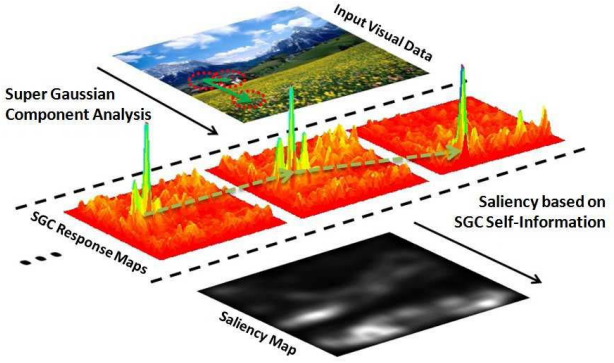

Fig. 1. What are we looking for when viewing a scene? Our studies suggest that the answer to this question could be revealed via statistical analysis of human eye fixations. One possible answer named super Gaussian component is investigated in this paper.

In our previous work [4] (Fig. 1), we constructed a generic statistical framework for both saccadic behavior simulation and visual saliency analysis. Differently with traditional works that drew inspirations from the existing neurobiological knowledge or mathematical theories, we directly make assumptions based on the statistical analysis of the ground truth human eye-fixations. By means of statistical analysis, we try to find out "what components in visual images draw fixations" which is similar but more reachable compared with the traditional question of "what properties draw attention". The analysis was conducted on eye fixation data captured from human observers using eye tracking devices during task independent free viewing of natural images. In such bottom-up scenario, we have found an interesting phenomenon, which was further proved, through dense experiments, as a general principle that stimuli with a super Gaussian distribution is more likely to gather human gaze. Based on this finding, human saccadic behavior was modeled as a function of active information pursuit targeting at the statistical components with desired properties such as super Gaussianity. We show an illustration of our statistical model in Fig. 2. In this framework, visual data is represented as an ensemble of small image patches. Kurtosis maximization is adopted to search for the Super Gaussian Component (**SGC**). A response map is then obtained by filtering the original image with the found SGC. Based on the response map, we adopt a well known principle named winner-takes-all (**WTA** [5]) to select and locate the simulated fixation point. Gram-Schmidt orthogonal method is applied at the

Manuscript received September 22, 2013; revised March 10, 2014 and June 17, 2014; accepted July 2, 2014. Date of publication July 10, 2014; date of current version September 23, 2014. This work was supported in part by the National Science Foundation of China under Grant 61071180 and in part by the Key Program Grant of National Science Foundation of China under Grant 61133003. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Carlo S. Regazzoni.

X. Sun, H. Yao, and X.-M. Liu are with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin 150001, China (e-mail: xiaoshuaisun@hit.edu.cn; h.yao@hit.edu.cn; xmliu@hit.edu.cn).

R. Ji is with the Department of Cognitive Science, School of Information Science and Engineering, Xiamen University, Xiamen 361005, China (e-mail: rrji@xmu.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/TIP.2014.2337758

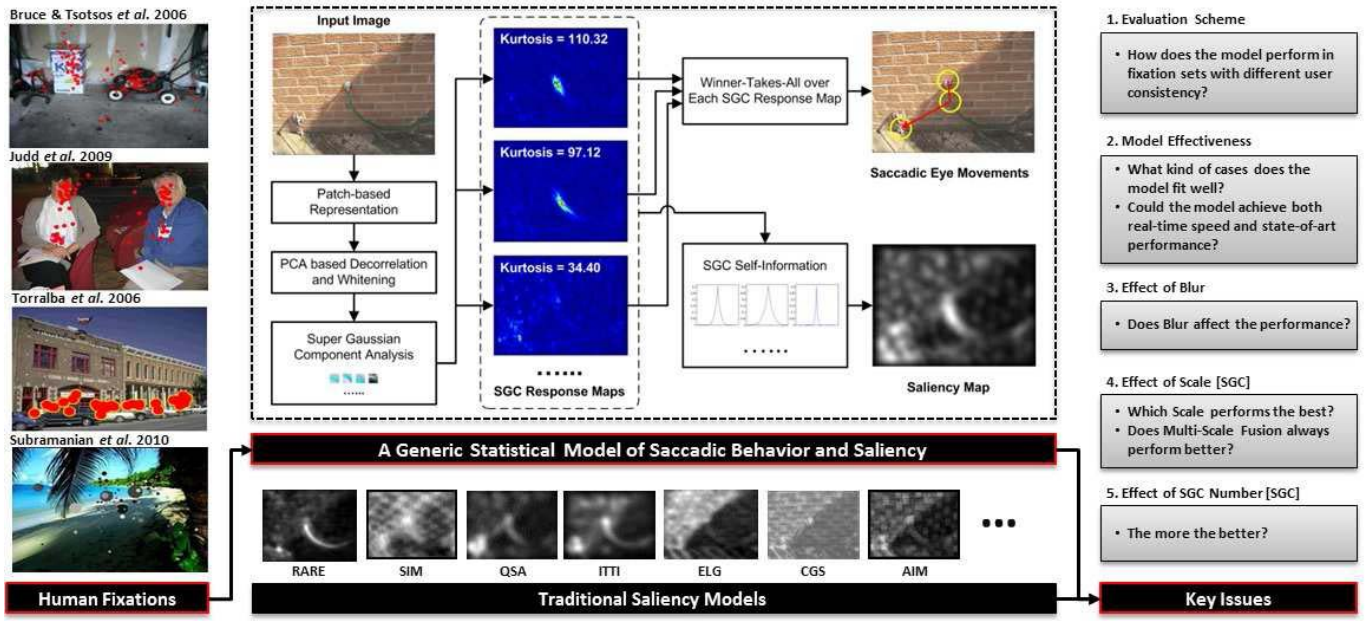

Fig. 2. The architecture of this paper. Firstly, we introduce more details about the analyzing results on human eye fixations as well as the implementation of our statistical model for saccadic behavior and bottom-up saliency. Then we present comprehensive experimental results, along with extensive comparisons and discussions on our model and 19 state-of-the-art approaches. Through these experimental studies, we try to unravel a series of inspiring issues of saliency modeling research, including 3 common issues for all saliency models and 2 specific issues for our **SGC** approach.

beginning of each selection to avoid convergence at the same location. Along with the saccadic simulation, a saliency map can also be estimated using either the selected fixations or the response maps. The proposed framework enables fast selection of a small number of fixations, which give processing priority to the most important components of the visual input. Different from low-level feature-based saliency driven approaches, the proposed gaze selection method is guided by high-level feature-independent statistical cues, which is supported by findings observed from real-world fixation analysis. Besides, the reliability and effectiveness of the **SGC** responses used in our framework have already been extensively validated in many statistical application such as Under-determined Blind Source Separation [6], [7].

The architecture of this paper is presented in Fig. 2. Firstly, we introduce more details about the analyzing results on human eye fixations as well as the implementation of our statistical model for saccadic behavior and bottom-up saliency. Then we present comprehensive experimental results, along with extensive comparisons and discussions on our model and 19 state-of-the-art approaches. Through these experimental studies, we try to unravel a series of inspiring issues of saliency modeling research, including 3 common issues for all saliency models (Fig. 2 issues 1-3) and 2 specific issues for our approach (Fig. 2 issues 4-5). Finally, we highlight the contribution of this paper by providing reasonable conclusions to all the above issues, which we believe will be instructive and inspiring for future research in this direction.

II. Related Works

In the literature, it is widely agreed that eye-movements are guided by both bottom-up (stimulus-driven) and top-down (task-driven) factors [2], [8], [9].

The bottom-up stimulus-driven research mainly focuses on saliency-driven approaches, in which a saliency map is pre-computed using low-level image features to guide task independent gaze allocation. These methods have been proven to be very effective in predicting eye fixations captured from human subjects while viewing natural images and video sequences. Itti *et al.* [2] proposed a computational attention model based on Koch and Ullman's attentional selection architecture [5], in which visual saliency is measured by spatial center-surround differences across several feature channels and different scales. In the model of [2] and [10], two principles named winner-takes-all (**WTA**) and inhibition-of-return (**IoR**) are adopted to select fixations based on saliency maps. This technique is widely used for scanning visual scene or generating artificial saccades. Bruce and Tsotsos [11] proposed a framework based on image sparse representation and the principle of information maximization, where visual saliency is measured by the self-information of the sparse coefficients. Also based on sparse coding, Hou *et al.* [12] argued that visual saliency should be dynamically measured by the incremental coding length of the sparse features. Wang *et al.* [13] adopt Site Entropy Rate as a saliency measure based on some well acknowledged biological facts with respect to both sparse coding and neuron activities in human vision system. Integrated with more biological factors, Wang *et al.* [14] extended their model to simulate saccadic scanpaths on natural images. Despite the above models, there are also many other works which present insightful saliency measures such as Bayesian Surprise [15], Center-Surround Discriminant Power [16], Spatially Weighted Dissimilarity [17], Image Signature [18], Rareness [19] *etc.*. Bayesian Surprise [15] is defined by relative entropy of the visual features' prior and posterior probability distribution, which was demonstrated

to be an important factor to attract human attention. Gao *et al.* [16] transferred saliency detection task to a binary classification problem, then proposed a discriminant center-surround hypothesis of saliency based on mutual information and obtained an optimal solution from the decision-theoretic perspective. Duan *et al.* [17] adopted PCA-based feature representation and introduce Center-Bias to compute the Spatially Weighted Dissimilarity between image patches as the final saliency. Image Signature [18] was a frequency-based image descriptor which spatially approximates the foreground (conspicuous contents) of an image. Without complex theories, Riche *et al.* [19] and Borji *et al.* [20] designed efficient saliency detection algorithms which takes the rareness (Rarity) of features as the major principle for bottom-up attention.

For top-down research, Yarbus *et al.* [21] designed an eye-tracking system and scene viewing experiments which initially revealed the role of particular instructions in determining where the subjects look. Following Yarbus's pioneering work, there are continued extensive studies of human saccadic behavior during different real-world tasks such as making a sandwich, fixing a cup of tea or learning and matching a shape [22]–[25]). Most studies indicate that eye-movements in top-down scenarios are probably made to collect task-relevant information [25]. Foulsham *et al.* [26] ask the participants to view color photographs of natural scenes in preparation for a memory test. Eye movements were recorded during the viewing and testing process. Analysis on these eye-tracking data indicates that saliency model work much better than random models but still may be missing out on sequential aspects of oculomotor control that could potentially predict fixation better than saliency alone. Similar results are observed in Tatler *et al.* [27], where static saliency is proven to be a good predictor for human gaze in scene-viewing but unlikely to generalize to other situations and a set of principles underlying eye guidance involving behavioral relevance, reward, uncertainty, priors *etc.*. are also studied trying interpret goal-driven gaze behavior. Although top-down guidance is not available in many computer vision applications. we can still draw inspirations from the top-down research, *eg*. the environmental effect and the sequential characteristics. Thus, we build our framework not only based on statistical factors but also considering the sequential aspects of human perception.

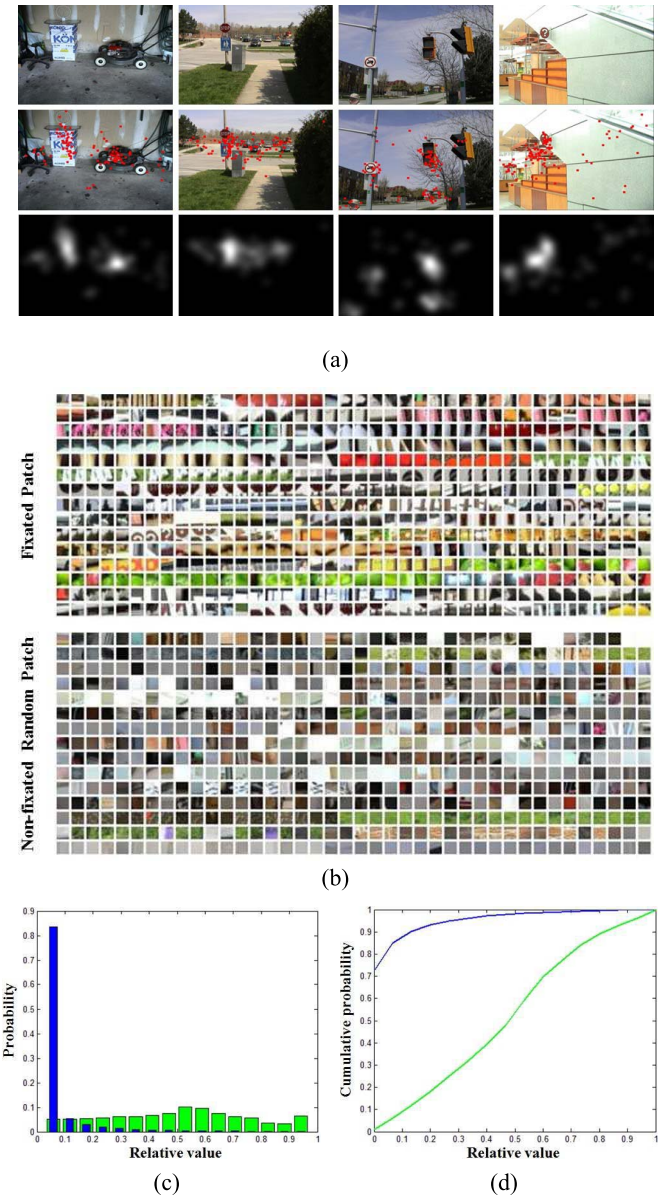

Fig. 3. Real-world distribution of bottom-up visual saliency. In (a), Top: natural images; Middle: images covered with human eye fixations (red dots); Bottom: eye-fixation density maps. (b) shows patch examples randomly sampled from fixated and non-fixated locations (Patches from the same image are grouped together); (c) presents the probability density of saliency (blue) and image pixels (green), where both quantity are normalized to [0, 1]; (d) is the corresponding cumulative distribution of (c).

III. Statistical Analysis of Human Fixation Data

Although there are many research works that address the problem of saliency detection, the statistical analysis of saliency is still non-trivial because there are no recordable "ground truth" saliency maps. In [11], a fixation density map is produced for each image based on human eye fixation points. The fixation density map comprises the probability of each pixel in the image being sampled by human observers based on their eye fixations. Taking fixation density as the approximation of saliency makes it possible for us to quantitatively analyze the statistical properties of saliency. Specifically, we use eye fixation data from two benchmark datasets (Bruce *et al*. [11] and Judd *et al*. [28]) for intuitive and statistical analysis of human saccades. Note that, such investigation is very similar to the previous work of Tatler *et al.* [29], in which four visual clues including contrast, orientation energy, and chromaticity were proposed and demonstrated to be quite effective in distinguishing the fixated and non-fixated visual stimuli in natural scenes. In comparison with [29], we conduct our investigations by considering not only the explicit visual clues such as color and orientation, but also the implicit statistical cues like the spatial allocation and density distributions of human fixations. Fig. 3(a) shows some example images along with the eye fixation points and the corresponding density maps. Assuming that the vision system is programmed to search for some inherent statistical components when viewing a visual scene, we might be able to discover some common properties between the underlying

components and the data segments located by ground truth eye fixations. Following this intuition, we gathered two ensembles of color image patches, one was constructed by sampling patches from eye fixations and the other from random locations. Examples of these two patch ensembles are presented in Fig. 3(b). It clearly shows that Fixation patch contains much more structural information compared with Random patches. Taking the fixation density as an approximation for ground truth saliency, we give a further comparison, in Fig. 3(c) and (d), on the probability density distribution between normalized saliency and pixel values. From intuitive and statistical observation, we found two interesting characteristics of visual saliency:

- Saliency is very *sparse*, which means the saliency of most locations is zero and only a small portion of the image has obvious high saliency value;
- High saliency value tends to be located surrounding the regions with abundant structural information.

According to feature integration theory [30], saliency is obtained by integration of multiple feature channels. Thus, features used for saliency detection should share similar statistical characteristics with saliency. From a statistical point of view, the above characteristics of saliency share great similarity with super-Gaussianity, which is synonymous with "sparse" and "structurized" in statistics. Considering the above issues, we proposed the primary assumption that Super Gaussian Components (**SGC**) of the scene are exactly what we are looking for during the viewing process.

IV. The Model

In this section, we present two major components of our model in details including sequential gaze selection and visual saliency estimation. The sequential aspects and the statistical assumption we made in Section II are both considered in our model. Different with previous bottom-up methods, the proposed model is developed based upon a statistical prior directly concluded from the analysis of human eye fixation data and it works without any training procedure which is also distinguished from traditional supervised methods.

A. Sequential Gaze Selection

There is one statistical technique named projection pursuit that shares a similar sequential selection behavior with saccadic eye-movement. Projection Pursuit was initially proposed to identify potentially meaningful data structures in high-dimensional data sets by searching for statistically "interesting" low-dimensional projections [31]. As a simple yet powerful tool, projection pursuit is widely used for high-dimensional data visualization and statistical component analysis, *eg*. ICA [32]. A linear projection from $\mathbf{R}^d$ to $\mathbf{R}^k$ is any linear map $\mathbf{W}$ (mostly $k \times d$ matrix of rank k):

$$\mathbf{D_p} = \mathbf{WD}, \mathbf{D} \in \mathbf{R}^d, \mathbf{D_p} \in \mathbf{R}^k, \tag{1}$$

where $\mathbf{D}$ is a d-dimensional random variable with distribution P, and $\mathbf{D_p}$ is a k-dimensional random variable with distribution P_W. Each column of $\mathbf{D_p}$ can be regarded as an unique feature (component) channel. Based on above definition, Projection Pursuit searches for a projection $\mathbf{W}$ maximizing (or minimizing) a certain objective function $G(P_W)$.

In previous works, Projection Pursuit were mostly used for searching interesting projections in which the data are separated into distinct, meaningful clusters. In the case of saccadic modeling, we adopt projection pursuit to search for the **SGC**s, which are further used for gaze localization and saliency estimation. From a signal processing point of view, this scheme can also be regarded as unsupervised function that dynamically separates the image data into a salient gaze-favored part and non-salient unattractive part. Technique details are described in IV-A1 and IV-A2.

1) Super Gaussian Component Analysis: Given an image I, we first turn it into a patch-based representation $\mathbf{X}$ by scanning I with a sliding window from top-left to bottom-right. $\mathbf{X}$ is stored as a $M \times N$ matrix, where each row vector corresponds to a reshaped RGB image patch (N is the number of patches, M the patch size). PCA based decorrelation and whitening are applied to $\mathbf{X}$ as a preliminary process, resulting in a new matrix $\mathbf{Z}$, which will simplify the subsequent calculations [32].

Single SGC pursuit - In statistics, the super-Gaussianity of a random variable is usually measured by the kurtosis function which is defined as:

$$\text{kurt}(y) = \mathbf{E}\{y^4\} - 3(\mathbf{E}\{y^2\})^2, \tag{2}$$

where y is the given random variable, $\mathbf{E}\{.\}$ is the expectation function. If y is a gaussian random variable, kurt(y) will be 0. If kurtosis is positive, the variable is called super-Gaussian which is also an alternative definition of sparsity. For whitened variable y, its standard deviation $\mathbf{E}\{y^2\} = 1$. So the kurtosis function can be further simplified as $\mathbf{E}\{y^4\} - 3$. To maximize the kurtosis, *ie*. maximizing super-Gaussianity, we can start from a random selected projection $\mathbf{w}$, and iteratively change its direction using fixed-point iteration method based on the available samples denoted as $\mathbf{Z}$. Now we give a formal objective function G_p for single SGC pursuit:

$$G_p(\mathbf{w}) = \text{kurt}(\mathbf{w^T Z}). \tag{3}$$

The gradient of G_p has the following form:

$$\frac{\partial G_p}{\partial \mathbf{w}} = 4[\mathbf{E}\{(\mathbf{w^T Z})^3 \mathbf{Z^T}\} - 3\mathbf{w^T} \|\mathbf{w^T}\|^2]. \tag{4}$$

During optimization, the iteration reaches convergence when the gradient vector and the projection vector have the same direction. Let Eq. 4 be equal to $\mathbf{w}$, we have:

$$\mathbf{w^T} \propto [\mathbf{E}\{(\mathbf{w^T Z})^3 \mathbf{Z^T}\} - 3\mathbf{w^T} \|\mathbf{w^T}\|^2]. \tag{5}$$

Eq. 5 leads to a fixed-point iteration algorithm:

$$\begin{aligned} \mathbf{w} &\leftarrow \mathbf{E}\{\mathbf{Z}(\mathbf{Z^T w})^3\}, \\ \mathbf{w} &= \mathbf{w}/\|\mathbf{w}\|. \end{aligned} \tag{6}$$

Based on Eq. 6, we can get a projection vector which maximizes the super-Gaussianity of the projected data. There are two conditions that will make the iteration stop:1. $\|\triangle \mathbf{w}\| < \epsilon$, where $\triangle \mathbf{w}$ is the difference of $\mathbf{w}$ after one iteration and ϵ is a convergence threshold; 2. Optimization didn't converge within a limited number of iterations.

Fig. 4. Eye-movements generated by our model. For each image we show the scanpath with five saccades and the corresponding focused regions.

Fig. 5. Comparisons of fixation density maps between the proposed model and human observers. Top: input images; middle: fixation density maps generated by our model using 75 fixations per image; bottom: fixation density maps of human eye fixations. The spatial distribution of our model-generated fixations are very similar with those of humans.

Multiple SGC pursuit - Based on the single component pursuit method, multiple SGC pursuit can be implemented by applying the same optimization method under the constrain that the new SGC should be orthogonal to the previous ones. The orthogonal constrain prevents the optimization process from converging on the same defections of the previous pursuit process. Practically, we use Gram-Schmidt orthogonal method for orthogonalization. Given a set of predefined projections: $\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_p$, we ensure the orthogonal constrain by adding the following orthogonalization procedure:

$$\mathbf{w}_{p+1} \leftarrow \mathbf{w}_{p+1} - \sum_{j=1}^{p} (\mathbf{w}_{p+1}^{\mathbf{T}} \mathbf{w}_j) \mathbf{w}_j . \tag{7}$$

The normalization $\mathbf{w} = \mathbf{w}/\|\mathbf{w}\|$ in Eq. 6 can be repositioned at the end of each iteration. Based on Eq. 6 and Eq. 7, multi-SGC pursuit can be performed in a one-by-one manner.

2) WTA Based Gaze Localization: For each super Gaussian component, we generate a response map by treating the component as a linear filter:

$$\mathbf{RM}_i = \mathbf{w}_i^{\mathbf{T}} \mathbf{Z} \tag{8}$$

where $\mathbf{RM}_i(j)$ denotes the response value of jth patch for the ith SGC. Similar with [2], we select the location with largest response value as the gaze point following the WTA principle. Fig. 4 and Fig. 6 shows the visualized gaze selection process on natural images, along with the focused local-regions. Statistical similarity of gaze selection behavior between human observers and the proposed model could be observed from the fixation density maps in Fig. 5. Note that, the orthogonal constrain forces the model not to attend to a SGC that has already been perceived. This could bring potential problems because there might be multiple locations that give strong responses to a single SGC, while our current scheme could only attend to the strongest one and leave the others out of focus. A quick solution to this weak point is to adopt spatial-level **IOR** within each SGC response map, which could spot multiple high response locations and also eliminate the potential conflict between multiple winners.

Algorithm 1 Gaze Selection and Saliency Estimation

Input: $M \times N$ data matrix $\mathbf{Z}$, $M \times M$ zero matrix $\mathbf{B}$, maximum iteration $\theta = 500$, convergence threshold $\epsilon = 0.0001$, maximum number of SGCs $N_C = 75$
Output: Fixation sequence F, Saliency map $\mathbf{S}$

Set projection index $k = 1$
while $k < Mork < N_C$ **do**
  Generate random vector $\mathbf{w} = [w_1, w_2, ..., w_M]$
  Orthogonalize $\mathbf{w}$ by $\mathbf{w} = \mathbf{w} - \mathbf{B}\mathbf{B}^{\mathbf{T}}\mathbf{w}$
  $\mathbf{w} = \mathbf{w}/\|\mathbf{w}\|, j = 1$
  while $j < \theta$ and $\|\mathbf{w}' - \mathbf{w}\| < \epsilon$ **do**
    $\mathbf{w}' = \mathbf{w}$
    $\mathbf{w} = \mathbf{Z}(\mathbf{Z}^{\mathbf{T}}\mathbf{w})^{\mathbf{3}}/N$
    $\mathbf{w} = \mathbf{w}/\|\mathbf{w}\|$
    $j = j + 1$
  end while
  Replace the kth column of $\mathbf{B}$ by $\mathbf{w}$
  $\mathbf{RM}_k = \mathbf{w}^{\mathbf{T}}\mathbf{Z}$
  $F = F \bigcup < k, \text{argmax}\mathbf{RM}_k >$
  $k = k + 1$
end while
Generate $\mathbf{S}$ based on Eq. 9
Normalize $\mathbf{S}$ by mapping $[\mathbf{mean}(\mathbf{S}), \mathbf{max}(\mathbf{S})]$ to [0,1] to eliminate potential noises.
Smooth $\mathbf{S}$ with a gaussian filter $(5 \times 5, \sigma = 2)$
return $F, \mathbf{S}$

B. Visual Saliency Estimation

We measure visual saliency by the self-information of the super Gaussian components of the image. As the SC components are acquired sequentially in our model, the saliency map is estimated also in a dynamical manner. The more SG components are involved, the more details will appear in the saliency map. Given k response maps: $\mathbf{RM}_1, \mathbf{RM}_2, \ldots, \mathbf{RM}_k$, the bottom-up saliency of jth patch is defined as:

$$\mathbf{S}(j) = -\log \prod_{i=1}^{k} p_i(\mathbf{RM}_i(j)) = -\sum_{i=1}^{k} \log p_i(\mathbf{RM}_i(j)), \tag{9}$$

where $p_i(.)$ is the probability density function of the i-th SG component, which can be estimated using all the image patches obtained from the input image. For simplicity, we estimate $p_i(.)$ by histogram method which makes the result also an approximation of feature rarity. The detailed implementation for both gaze selection and saliency estimation is presented in Alg. 1 with default parameter settings.

C. Multi-Scale Processing

By changing the size of the sliding window, we can obtain multiple saliency maps focusing on different salient patterns at various scales. Thus a multi-scale saliency map can be obtained by fusing the saliency responses from all single-scale saliency maps: $\mathbf{S}_m(j) = \sum_{i \in \phi} \mathbf{S}_i(j)$, where $\phi = \{1, 3, 5, 7\}$ and $\mathbf{S}_i$ denotes the saliency map computed using image

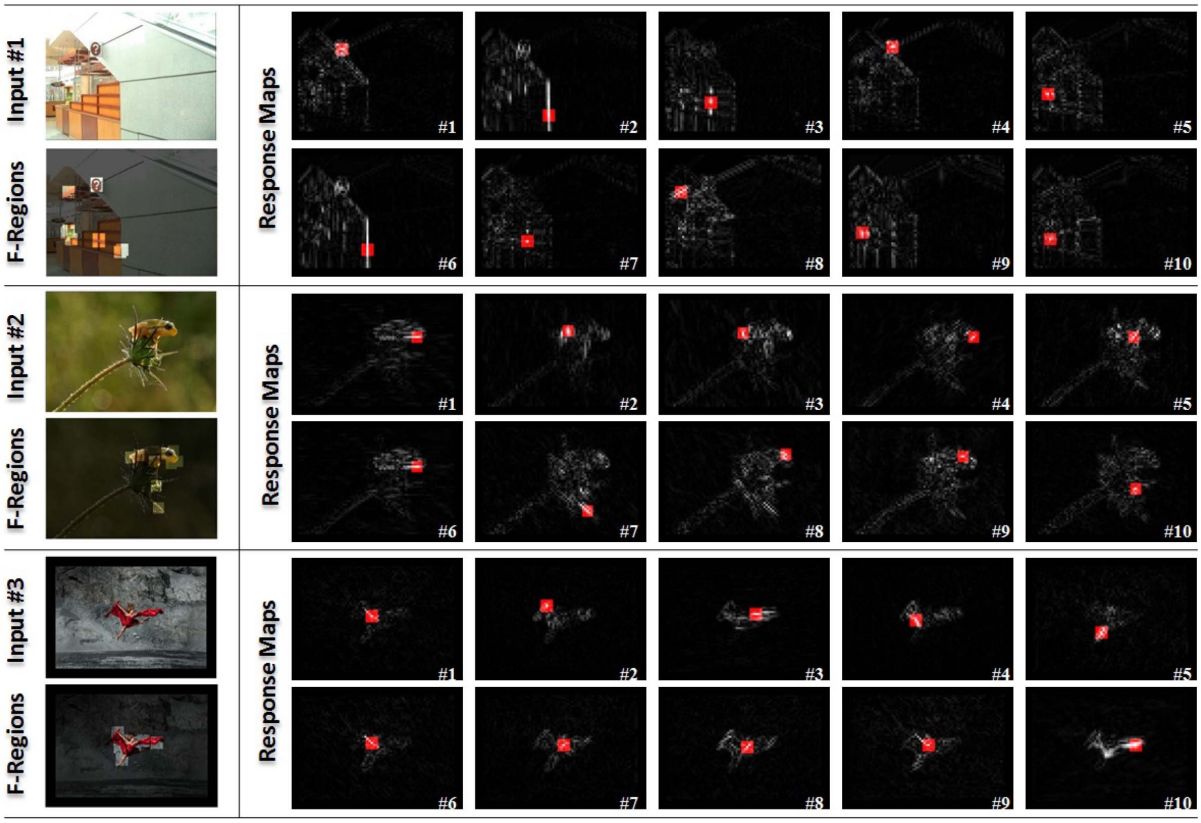

Fig. 6. Super Gaussian component analysis and gaze selection on natural images. In each group of results, we present the original image, the response maps of the top ten super-Gaussian components, the locations of gaze (red boxes), and the highlighted fixated-regions.

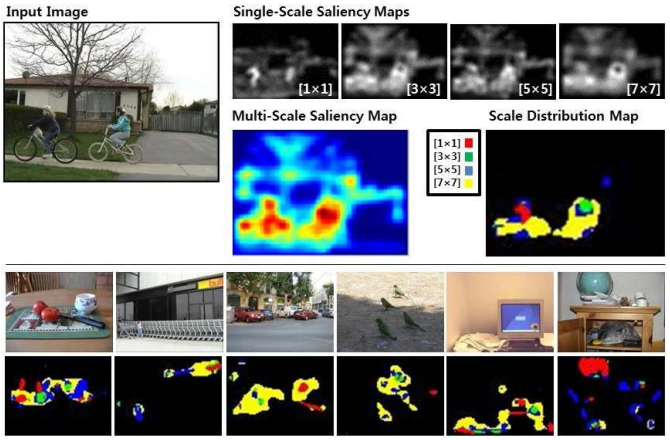

Fig. 7. Examples of our single / multi-scale saliency estimation. The scale distribution map adopts four colors to show the scale index with the maximum saliency value at a given location. The presented results demonstrate that our method are able to identify salient contents at different scales.

patches sampled by an $[i \times i]$ sliding window. Examples of both single and multi-scale saliency estimation results are presented in Fig. 7. To ensure the evaluation efficiency, the inputs are resized to have a max side length of no more than 80 pixel and it takes about 9s for the model to generate a multi-scale saliency map (Intel Core i5 3.1GHz, 12GB RAM). Quantitative experimental results indicate that our multi-scale fusion scheme can achieve significant improvements over the best single-scale strategy ($i = 5$) which has already outperformed most of the state-of-the-arts.

V. Experiments

In this section, we evaluate our model from two major aspects: 1) Qualitative experiments including the responses to psychological patterns, saliency-preserving image retargeting and robustness against visual distortions; 2) Quantitative evaluation of the performance in predicting human eye-fixations and detecting proto-object.

A. Response to Psychological Patterns

Responses to psychological patterns adopted in attention related experiments can indicate the biological plausibility of the tested models. This experiment was extensively studied in previous works of Bruce *et al.* [33] and Guo *et al.* [34], so we mainly compared our results with their models. For each image, we present the saliency maps generated by the three models and use red circles to indicate the locations of the maximum value in each saliency map. As shown in Fig. 8, all three models generate reasonable responses for normal salient patterns including density, orientation, color, curve, insertion and inverse-intersection. The Differences emerge in patterns with conjunctive features. In Fig. 8(b), we show three images with salient singletons, where the first one is defined by color, the second by orientation and third by both color and orientation. In this case, our model successfully discovered the most salient patterns under the conjunction of both color and orientation features, while the other two models are obviously more sensitive to orientation features.

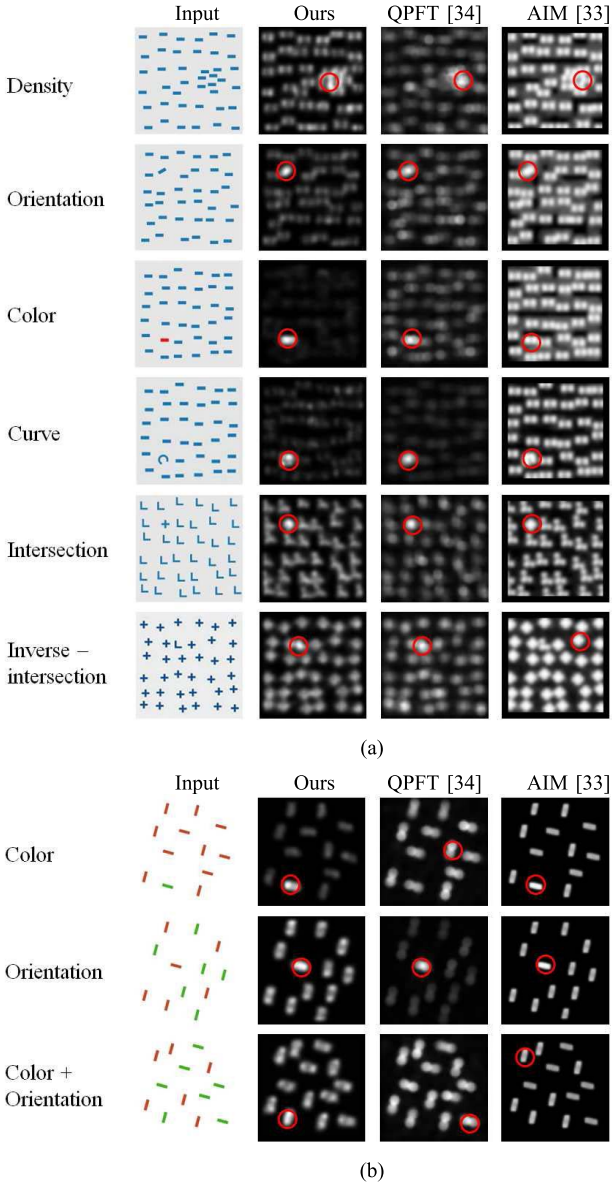

Fig. 8. Response to psychological patterns. From left to right, we present the input image, saliency maps generated by our method, QPFT [34] and AIM [33]. Red circles indicate the location of the maximum saliency. (a) Patterns with single salient feature. (b) Patterns with salient singletons.

The results reveal two important characteristics of our model: 1) It is able to capture outliers of feature dimensions that are not explicitly modeled, *eg*. it detects obviously orientation outliers, curvature and intersections without computing gradients etc. preliminarily. This is because our component pursuit strategy represents the image patches adaptively according the pattern statistics within the visual context specified by the input image, and such representation can highlight the desired statistical property, e.g. super Gaussianity, naturally and effectively for the subsequent saliency computation; 2) It is able to detect the salient patterns with conjunctive features which is somehow inconsistent with the perception characteristic of human vision. Humans do not find such outliers quickly, but perform serial search [30] instead. From technical point of view, this could be a good point, since we get an algorithm that can efficiently search for the super Gaussian components and outperform humans in finding salient patterns. On the other hand, this can also be a weak point, because such results did not fit human vision well.

B. Robustness Test

We test our model with manually modified images which contain commonly encountered visual distortions. As demonstrated in Fig. 9, our model is basically not influenced by various distortions including salt noise, gaussian noise, brightness & contrast change and down sampling. In addition, the most salient region indicated by our saliency maps remain basically the same under all types of distortions.

C. Saliency-Preserving Image Retargeting

The aim of automatic image retargeting is to resize an image by expanding or shrinking the non-informative/unimportant regions [35]. Since saliency can effectively distinguish the important visual contents from the cluttered background, it can be directly applied in the resizing process. Here, we combined our saliency model with a well acknowledged retargeting algorithm, namely Seam Carving (SC [35]), to generate saliency-preserving retargeted images. The SC algorithm functions by constructing a number of seams (paths of least importance) in an image and automatically removes seams to shrink the image or inserts seams to expand it. We modified the original code of [35] making the saliency map as an additional weight matrix for the seam carving algorithm. Typical results are presented in Fig. 10. The guidance of our statistical saliency successfully prevented the retargeting algorithm from damaging important visual contents such as faces and objects. Results generated using saliency maps of other competitive saliency models are also presented. In most of the cases, **RARE** [19] and our model introduce much less deformations and performed relatively better than the other methods.

D. Human Eye-Fixation Prediction

This experiment is designed for evaluating the consistency between human eye fixations and the saliency map generated by the tested models. Experiments are conducted on three data sets: static image data from Bruce *et al.* [11],[1] Judd *et al.* [28][2] and dynamic video data from Itti *et al.* [15].[3] Models listed for comparisons are selected by 2 criterions: 1. commonly used benchmark and 2. open source. Following the proposal of [11], [15], and [36], we adopt area under the ROC curve (**AUC**) and K-L divergence for quantitative evaluation. Due to the random factors and interpersonal differences, the sequences of saccades are hard to compare in this work as well as in most state-of-the-arts [2], [12], [13], [28], [33], [36]–[38]. The saccadic movements might be completely different even though they are captured from the same subject when viewing the same scene. Therefore, to guarantee the generality and fairness of the comparison, we applied **AUC** and K-L divergence as the main evaluation metric for eye fixation prediction,

[1] http://www-sop.inria.fr/members/Neil.Bruce/
[2] http://people.csail.mit.edu/tjudd/WherePeopleLook/
[3] http://crcns.org/data-sets/eye

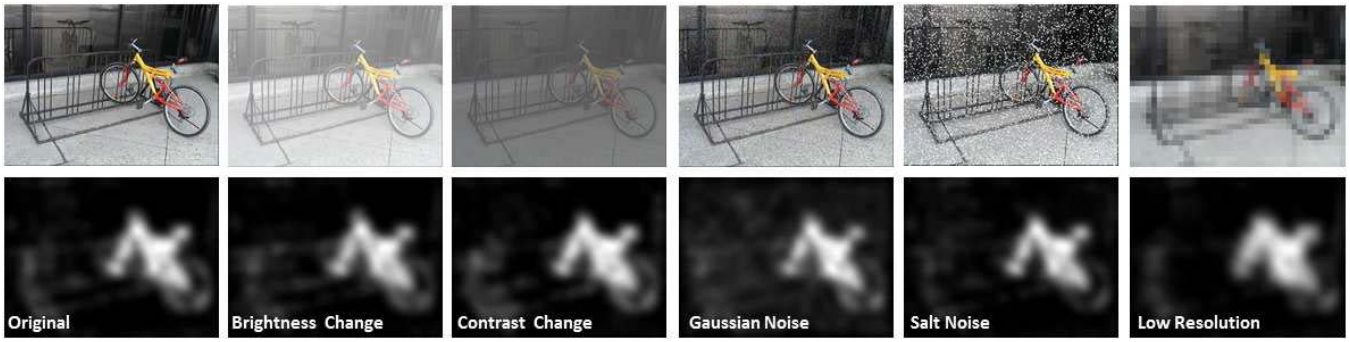

Fig. 9. Our model is robust to various distortions including Contrast & Brightness change, Salt & Gaussian noise and extremely Low Resolution [30 × 40].

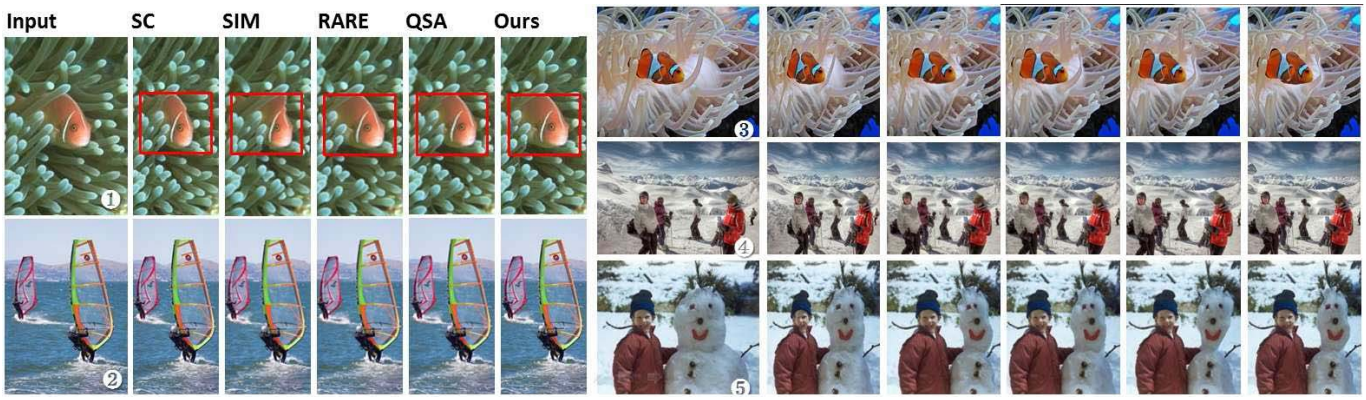

Fig. 10. Examples of Saliency-Preserving Image Retargeting. In each group, we present the original input, the resized images by seam carving [35], our method and other competitive saliency models. 30% of the seams are removed from the inputs.

which is also reliable and commonly used in computer vision community [13], [39], [40].

The original **AUC&KL** evaluation is largely affected by the "edge effect" due to center bias caused by the central composition of interesting objects. Zhang *et al.* [36] pointed out that a simple gaussian blob fitted to the eye fixations has a **AUC** score of 0.80 which exceeds most of the reported models on Bruce's data set [11]. To eliminate the interference caused by the "edge effect", we follow the proposal of Zhang *et al.* [36], and use a refined evaluation procedure to compute the **AUC** score. Specifically, we first compute the true positives from the saliency maps based on the human eye fixation points. In order to calculate false positives, we use the human fixation points from other images by permuting the order of images. This permutation of images is repeated for 100 times. Each time, we compute the **AUC** and **KL** score by regarding the eye fixations from original image as the positive samples and the fixations from permutated images as the negative samples. The KL-Divergence between the distribution of two sample sets is computed using 16-bin histograms.

According to the collection scope of fixation samples, the evaluation metrics can be further divided into two levels: Set-Level and Image-Level. The Set-Level metrics collect fixation samples from the entire database. It shows the generic performance of tested models over datasets; The Image-Level metrics compute the AUC and KL scores using samples from a single image, which allows us to check the specific performance of models over individual images. In summary, we have 4 evaluation metrics including Set-Level AUC (**SL-AUC**), Set-Level KL (**SL-KL**), Image-Level AUC (**IL-AUC**) and Image-Level KL (**IL-KL**).

1) Experiment on Static Natural Images: Fixation dataset from Bruce and Tsotsos [11] (YORK-120) contains 11,999 eye fixations captured from 20 human subjects free viewing 120 natural images for 4 seconds each. To reduce the influence caused by the subjects' personalized factors, we create two sub-dataset (YORK-120-SUB-1 & 2) by filtering out spatially isolated saccades using the fixation density maps which are also included in the dataset package. Each fixation density map is normalized to [0, 1]. YORK-120-SUB-1 is consist of 8,190 fixations with density value greater than 0.2 and YORK-120-SUB-2 contains 4,339 fixations with density value greater than 0.5. The MIT-1003 [28] dataset contains 1,003 images from Flickr and LabelMe covering 779 landscape scenes and 228 portrait photos. To ensure fair comparisons to those supervised approaches, we randomly picked 100 images, as suggested in [28], for each round of the evaluation. Again, to reduce the interferences caused by interpersonal differences, we only keep the first 6 saccades per subject to construct the testing dataset resulting in a total of 19,812 fixations.

We compared our model against 19 state-of-the-art approaches including **ITTI** (Itti *et al.* [2]), **ICL** (Hou *et al.* [12]), **SER** (Wang *et al.* [13]), **AIM** (Bruce and Tsotsos [33]), **SR** (Hou *et al.* [37]), **QPFT** (Guo *et al.* [34]), **SUN** (Zhang *et al.* [36]), **SIM** (Murray *et al.* [39]), **JUDD** (Judd *et al.* [28]), **CGS**

TABLE I
EYE FIXATION PREDICTION RESULTS I ($D_1 \sim D_3$ DENOTE YORK-120-ALL, YORK-120-SUB-1 AND YORK-120-SUB-2)

	SL-AUC			SL-KL			IL-AUC			IL-KL		
	D_1	D_2	D_3	D_1	D_2	D_3	D_1	D_2	D_3	D_1	D_2	D_3
AIM [33]	0.6700	0.7091	0.7464	0.1941	0.3117	0.4566	0.6767	0.7166	0.7608	1.1270	2.3390	4.3290
CAS [45]	0.6830	0.7252	0.7759	0.2180	0.3461	0.5631	0.6906	0.7425	0.7873	1.3680	3.0280	5.8210
CGS [41]	0.6648	0.6934	0.7269	0.1871	0.2696	0.4553	0.6603	0.6906	0.7272	0.9492	1.8470	3.5450
ELG [20]	0.6858	0.7215	0.7730	0.2374	0.3559	0.7177	0.6879	0.7304	0.7789	1.2330	2.5940	4.9690
FT [42]	0.5390	0.5491	0.5550	0.0122	0.0193	0.0301	0.5400	0.5569	0.5612	0.6747	1.1700	2.4140
GBVS [43]	0.6372	0.6839	0.7422	0.1242	0.2254	0.4248	0.6374	0.6943	0.7484	1.2700	3.0270	5.8660
HFT [44]	0.6668	0.7146	0.7651	0.1820	0.3433	0.5862	0.6710	0.7233	0.7754	1.3230	3.0600	5.9210
ICL [12]	0.6781	0.7203	0.7724	0.2120	0.3472	0.6040	0.6955	0.7521	0.8097	1.3650	3.1140	5.9790
ITTI [2]	0.5932	0.6141	0.6473	0.0593	0.0872	0.1497	0.5926	0.6218	0.6531	0.6736	1.0830	1.9060
QPFT [34]	0.6880	0.7317	0.7874	0.2358	0.3775	0.6645	0.7041	0.7533	0.8094	1.2580	2.5540	5.1310
QSA [46]	0.6946	0.7384	0.7834	0.2566	0.4352	**0.9149**	0.7044	0.7540	0.8032	1.3660	3.1200	6.0280
RARE [19]	0.6957	0.7423	0.7969	0.2583	0.4250	0.6979	0.7040	0.7562	0.8104	1.4050	3.1950	6.0020
SER [13]	0.6613	0.7046	0.7599	0.1741	0.2953	0.5181	0.6845	0.7452	0.7993	1.3450	3.1160	5.8270
SIG [18]	0.6260	0.6579	0.6947	0.0954	0.1519	0.2632	0.6428	0.6782	0.7209	0.5066	0.8847	1.7390
SIM [39]	0.7040	0.7434	0.7908	0.2870	0.4357	0.6805	0.7057	0.7544	0.8061	1.4660	3.1570	5.7880
SR [37]	0.6775	0.7148	0.7541	0.2034	0.3203	0.5017	0.6808	0.7217	0.7617	1.0980	2.2260	4.5820
STSR [40]	0.6925	0.7340	0.7865	0.2631	0.4328	0.8953	0.7001	0.7467	0.7999	1.5030	3.2590	6.0460
SUN [36]	0.6658	0.6997	0.7526	0.1868	0.2820	0.5516	0.6715	0.7131	0.7637	1.2260	2.4920	4.8800
JUDD [28]	0.6165	0.6493	0.6903	0.0931	0.1706	0.3226	0.6176	0.6578	0.7043	0.9830	2.2680	4.3770
Ours-SS	0.7100	0.7509	0.7994	0.2980	0.4543	0.7381	0.7104	0.7590	0.8058	1.4380	3.2010	5.8410
Ours-MS	**0.7153**	**0.7572**	**0.8099**	**0.3141**	**0.4768**	0.8959	**0.7168**	**0.7658**	**0.8148**	**1.5190**	**3.3640**	**6.0770**

TABLE II
EYE FIXATION PREDICTION RESULTS II ($D_4 \sim D_6$ DENOTE MIT-1003-ROUND-1, MIT-1003-ROUND-2, AND MIT-1003-ROUND-3)

	SL-AUC)			SL-KL			IL-AUC			IL-KL		
	D_4	D_5	D_6	D_4	D_5	D_6	D_4	D_5	D_6	D_4	D_5	D_6
AIM [33]	0.6519	0.6236	0.6292	0.1711	0.1182	0.1281	0.6574	0.6329	0.6357	1.6870	1.4660	1.5420
CAS [45]	0.6638	0.6367	0.6544	0.1946	0.1287	0.1759	0.6698	0.6447	0.6623	1.9860	1.7250	1.7450
CGS [41]	0.6516	0.6259	0.6439	0.2168	0.1324	0.1998	0.6530	0.6353	0.6400	1.2210	1.0050	1.1140
ELG [20]	0.6671	0.6456	0.6505	0.1950	0.1429	0.1812	0.6689	0.6550	0.6565	1.7530	1.5100	1.6180
FT [42]	0.5366	0.5222	0.5117	0.0150	0.0112	0.0065	0.5397	0.5179	0.5083	1.0290	0.9518	0.9448
GBVS [43]	0.6181	0.5909	0.5964	0.0954	0.0625	0.0714	0.6263	0.5981	0.6006	1.8280	1.7300	1.7600
HFT [44]	0.6417	0.6153	0.6192	0.1383	0.0938	0.1065	0.6458	0.6198	0.6208	2.0010	1.6790	1.7930
ICL [12]	0.6503	0.6293	0.6423	0.1573	0.1221	0.1534	0.6474	0.6487	0.6580	2.0000	1.8160	1.8670
ITTI [2]	0.5736	0.5626	0.5523	0.0476	0.0393	0.0278	0.5752	0.5640	0.5519	0.8221	0.7575	0.7334
QPFT [34]	0.6528	0.6293	0.6373	0.1577	0.1111	0.1362	0.6597	0.6537	0.6564	1.5440	1.3620	1.4040
QSA [46]	0.6627	0.6394	0.6461	0.1892	0.1357	0.1585	0.6661	0.6493	0.6511	1.8190	1.6060	1.6650
RARE [19]	0.6807	0.6563	0.6740	0.2246	0.1709	0.2084	0.6759	0.6577	0.6819	1.8840	1.8030	1.9020
SER [13]	0.6354	0.6166	0.6244	0.1323	0.0963	0.1171	0.6381	0.6405	0.6414	1.9830	1.7780	1.7270
SIG [18]	0.6067	0.5979	0.5999	0.0650	0.0554	0.0589	0.6178	0.6073	0.6138	0.5337	0.5350	0.5361
SIM [39]	0.6704	0.6432	0.6503	0.2025	0.1441	0.1673	0.6732	0.6515	0.6578	1.9350	1.7540	1.7810
SR [37]	0.6507	0.6262	0.6333	0.1544	0.1070	0.1324	0.6617	0.6424	0.6349	1.4210	1.2420	1.2540
STSR [40]	0.6660	0.6263	0.6400	0.2078	0.1230	0.1644	0.6562	0.6229	0.6455	2.0200	1.7640	1.8180
SUN [36]	0.6391	0.6275	0.6419	0.1493	0.1197	0.1589	0.6363	0.6385	0.6478	1.7870	1.6280	1.7750
JUDD [28]	0.6102	0.5872	0.6026	0.0957	0.0626	0.0871	0.6112	0.5837	0.6026	1.4140	1.3190	1.2600
Ours-SS	0.6864	0.6638	0.6781	0.2464	0.1923	0.2357	0.6760	0.6655	0.6777	2.0790	1.8960	1.9350
Ours-MS	**0.6934**	**0.6705**	**0.6881**	**0.2619**	**0.2082**	**0.2634**	**0.6817**	**0.6701**	**0.6888**	**2.1530**	**1.9720**	**2.0510**

(Torralba *et al.* [41]), **ELG** (Borji *et al.* [20]), **FT** (Achanta *et al.* [42]), **GBVS** (Harel *et al.* [43]), **HFT** (Li *et al.* [44]), **CAS** (Goferman *et al.* [45]), **QSA** (Schauerte *et al.* [46]), **RARE** (Riche *et al.* [19]), **SIG** (Hou *et al.* [18]), and **STSR** (Seo *et al.* [40]). Saliency maps of all tested methods are generated using their default parameter settings. The full results are presented in Table I and Table II, where $D_1 \sim D_6$ denote YORK-120-ALL, YORK-120-SUB-1, YORK-120-SUB-2, MIT-1003-ROUND-1, MIT-1003-ROUND-2, and MIT-1003-ROUND-3 respectively. Our method outperforms most of the state-of-the-art models in all 4 evaluation metrics. Fig. 11 compares **SL-AUC** and **SL-KL** scores of the tested models more intuitively. We show more visual comparisons in Fig. 12.

2) Experiment on Dynamic Videos: Video processing is slightly different from static image processing: 1. videos are processed into spatio-temporal patches [3 × 3 × 3 × 3] (Height × Width × Color × Frame) which are also reshaped into 1-D vectors during the Super-Gaussian Component Analysis stage; 2. to compare the saliency of the same visual content over time, we use Eq. 10 to normalize the saliency maps.

$$\mathbf{S}(x) = \delta(\eta \mathbf{S}(x)/\bar{\mathbf{S}}), \delta(x) = \begin{cases} 1 & \text{If } x > 1 \\ x & \text{else} \end{cases} \tag{10}$$

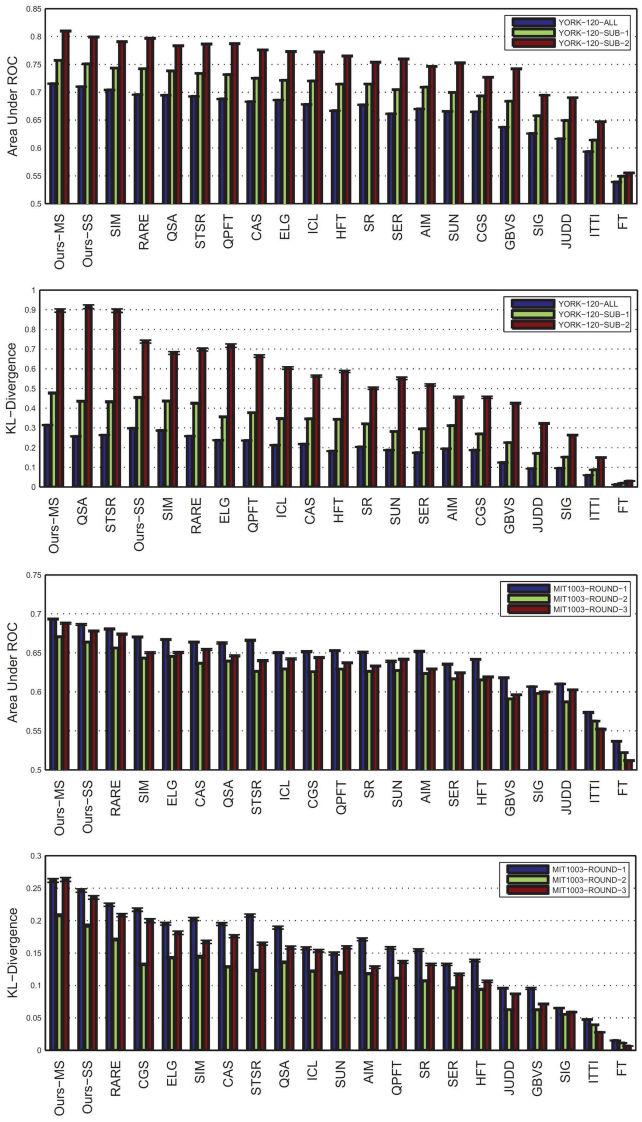

Fig. 11. SL-AUC & SL-KL results on YORK120 [11] and MIT 1003 [28]. The tested models are ranked according to their average performance under the corresponding evaluation metric. Our Single-Scale implementation (Ours-SS) outperformed most of the state-of-the-arts, while the Multi-Scale scheme (Ours-MS) achieved the overall best performance against all the tested models.

where $\bar{\mathbf{S}}$ is the mean value of $\mathbf{S}$, $\eta = 0.3$ is a scale parameter and fixed in the following experiment. Eye tracking data from Itti *et al.* [15] are recorded from 8 human subjects aged at 23-32 with normal vision. 50 video clips consisting various categories of dynamic scenes, including outdoor scenes, television broadcast and video games, are used for constructing the data set. 7 video clips of Berkeley outdoor scene consisting of 568 saccade points are used for evaluation. All results were produced using the attached evaluation scripts provided by [15]. As shown in Fig. 13, the **KL** score produced by our model is 0.692 ± 0.053 which is better than 0.530 ± 0.045 for Itti's saliency [2] and 0.589 ± 0.045 for surprise [15]. The **AUC** score of our model is 0.803 ± 0.009 which also outperforms the other two models (0.775 ± 0.011 for Itti's saliency, 0.776 ± 0.010 for surprise). We also compared our model with two static models, **RARE** [19] and **SIM** [39], which have been proved to be strong competitors in the image test. Although not specially designed for video signals, **RARE** also achieved remarkable results in the dynamic scene.

E. Proto-Object Detection

A candidate that has been detected but not yet identified as an object is defined as a *proto-object* [47]. In this experiment, we test the model's ability of detecting proto-objects in unconstrained natural scenes. The image data set, human label masks and evaluation codes used for this experiment are provided by Hou *et al.* [37]. Hit Rate (HR) and False Alarm Rate (FAR) are used for evaluating the saliency maps. Higher HR and lower FAR imply better detection performance. Quantitative comparisons between different saliency models are shown in Table III. We give two groups of results, one with fixed HR and the other with fixed FAR. Our model provides an overall better performance compared to Hou *et al.* [37], Seo *et al*. [40], Murray *et al*. [39] and Torralba *et al*. [41]. Fig. 14 shows more visual examples of our detection results.

VI. Discussions

A. Evaluation Scheme

Traditional evaluation schemes on human eye-fixation benchmarks use all the fixation samples in the dataset for evaluation. As discussed in [28], for some images, all viewers fixate on the same locations, while in other images viewers' fixations are dispersed all over the image. In our experiments, we characterize such inter-viewer consistency by separating the original dataset into several subsets with different level of fixation consistency (Sec. V-D1). As shown in Fig. 11, the results on YORK-120 dataset indicate that some models, *eg.***QSA** [46], **STSR** [40] and ours, are more powerful in distinguishing "commonly attended" salient regions from natural backgrounds. With this new evaluation scheme, we might be able to discover more insightful facts that could hardly be unraveled under the traditional schemes.

B. Model Effectiveness

What kind of cases does the model fit well? To answer this question, we conduct image level analysis based on the **IL-AUC** measure. In Fig. 15, we present the hot maps that show the **IL-AUC** scores for each image and each permutation during the evaluation on YORK-120 dataset. Given a hot map, each row presents the results of all permutations for a single image, each column shows the results of all images within a single permutation. By comparing the results from 8 competitive models, we could observe obvious similarity between the models in their image-wise performance. At the bottom, we show the well-handled and badly-handled images for **RARE** [19], **SIM** [39], **QSA** [46] and our model. The common good and bad cases also indicate the above-mentioned similarity among those models. From the fixation density maps presented in our case, we could see that the proposed SGC approach works better in images with limited salient objects and worse in cases with complex mixture of many objects. Such results might have two causes: 1) Using kurtosis as a measure of super-Gaussianity makes our

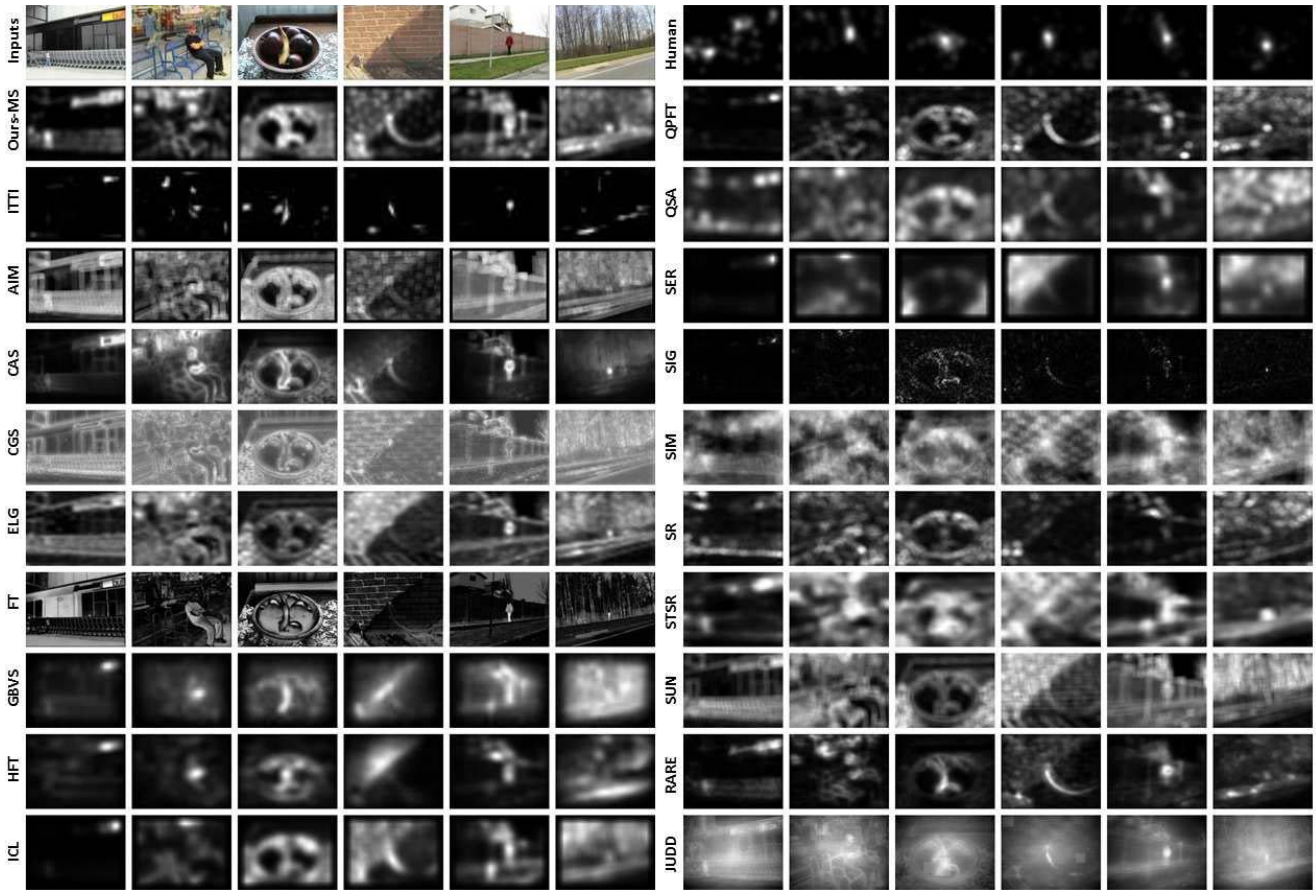

Fig. 12. Visual comparisons of our multi-scale saliency detection approach (Ours-MS) and 19 state-of-the-art saliency models on natural images.

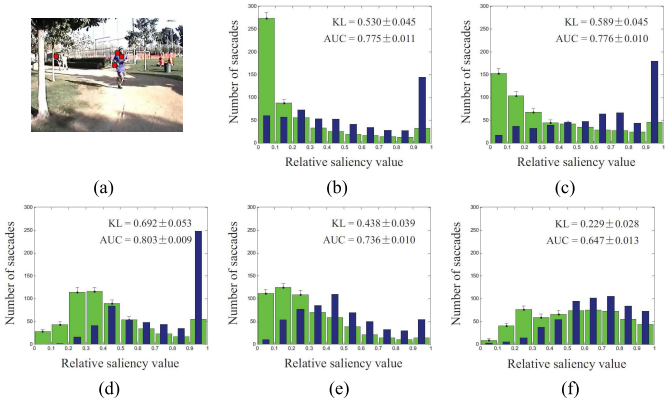

Fig. 13. (a) Input Example. (b) Saliency [2]. (c) Surprise [15]. (d) Our Model. (e) RARE [19]. (f) SIM [39]. Experimental results on dynamic scenes. 7 video clips of Berkeley outdoor scene [15] consisting of 568 saccade points are used for evaluation.

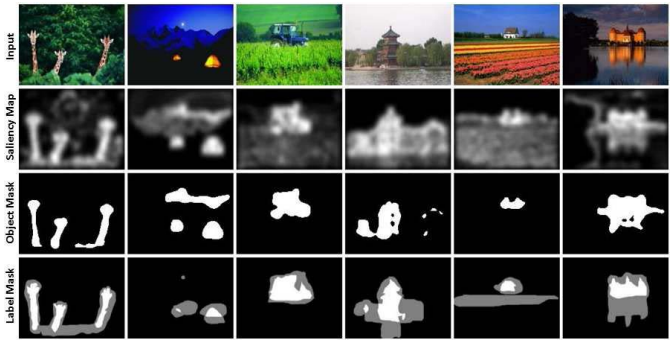

Fig. 14. Examples of saliency-based proto-object detection. In each column, we show the input image, our saliency map, binary object masks and human label masks. The object mask was obtained by binarizing the corresponding saliency map using an adaptive threshold $\alpha * \bar{\mathbf{S}}$. Here we set $\alpha = 2.5$. In each label mask, white color means the location is labeled as proto-object by all labelers, while gray means the location is labeled by at least one labeler.

object function sensitive to heavy tails and the multi-modality of functions [48], [49]; 2) It is not clear in human vision what the "correct" viewing behavior is. Besides, complex scenes could probably bring top-down interferences which will make the inter-viewer consistency of human fixations relatively low and thus harder to be reliably predicted. We also evaluated the time cost of the top-ranked saliency models on a machine with Intel Core i5 3.1GHz, and 12GB RAM. The average processing time of **RARE**, **SIM** and **QSA** for a single image are 1.21s, 4.26s and 0.029s respectively. For our model, the multi-scale approach cost 9s per image, and a simplified single scale version ([3 × 3], 1 SGC) only cost 0.03s but still maintained state-of-the-art performance (see Fig. 18 for more details).

C. Effect of Blur

Borji *et al.* [20] pointed out that blurring the saliency maps could affect the performance of the tested models under both traditional AUC and their proposed Shuffled AUC metric. In this section, we manually blur and evaluate the saliency maps of 4 top-ranked models in Sec. V with different size of

TABLE III

EVALUATION RESULTS OF PROTO-OBJECT DETECTION

	OURS Single-Scale / Multi-Scale	STSR Seo *et al.* [40]	SR Hou *et al.* [37]	SIM Murray *et al.* [39]	RARE Riche *et al.* [19]	QSA Schauerte *et al.* [46]	CGS Torralba *et al.* [41]
HR Fixed FAR	0.7495 / 0.8127 0.1433 / 0.1433	0.5933 0.1433	0.4309 0.1433	0.4878 0.1433	0.8342 0.1433	0.7468 0.1433	0.3712 0.1433
Fixed HR FAR	0.5076 / 0.5076 0.0867 / 0.0695	0.5076 0.1048	0.5076 0.1688	0.5076 0.1492	0.5076 0.0462	0.5076 0.0652	0.5076 0.1965

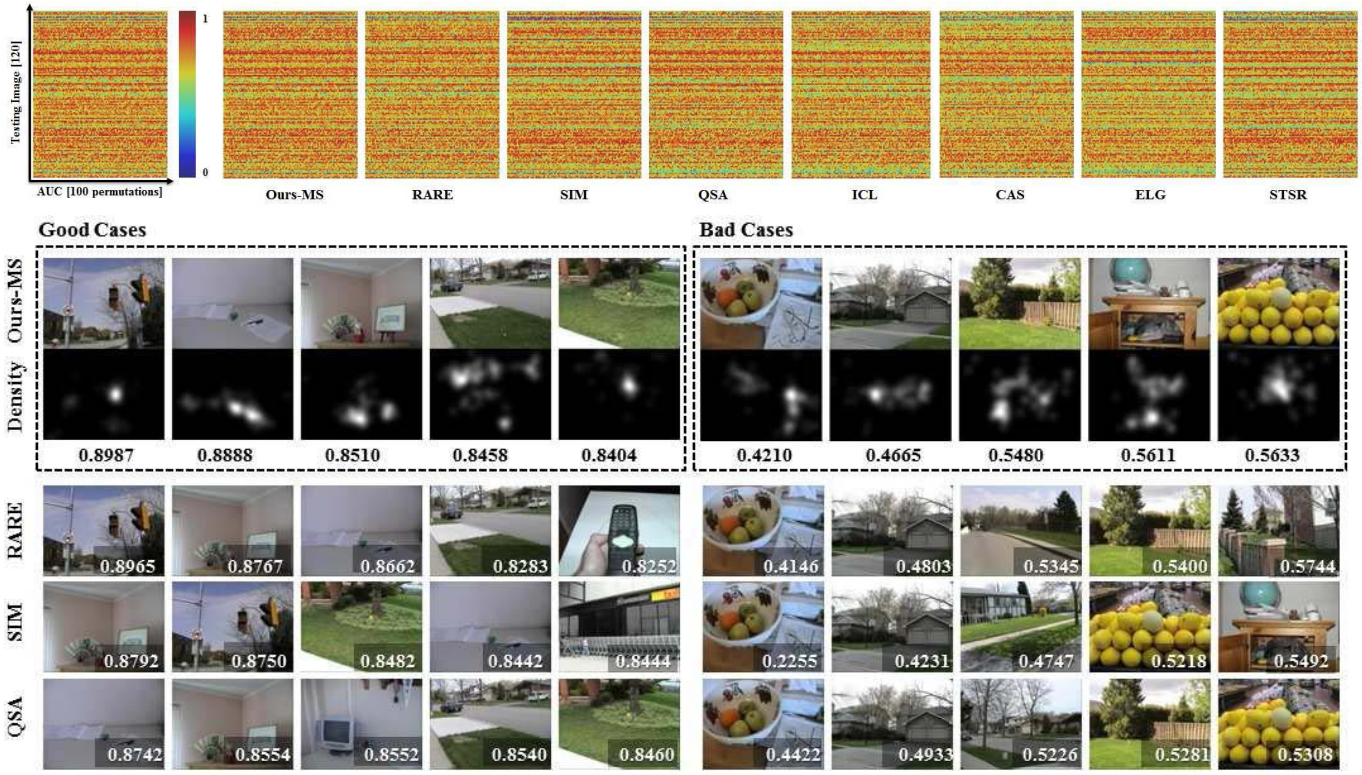

Fig. 15. Image level analysis on **IL-AUC** performance over the tested saliency models. Hot maps presented in the top row show the IL-AUC scores for each image and each permutation during the evaluation on YORK-120 dataset. Given a hot map, each row presents the results of all permutations for a single image, each column shows the results of all images within a single permutation. At the bottom, we show the well-handled and badly-handled images for RARE [19], SIM [39], QSA [46] and our model. In our case, we also present the density maps of the human fixations.

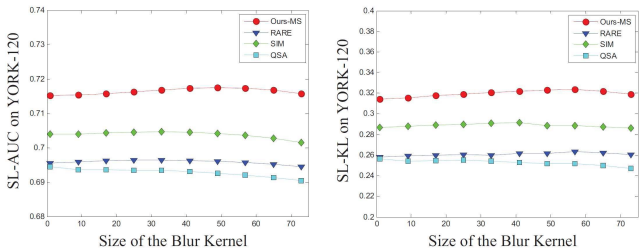

Fig. 16. SL-AUC (left) and SL-KL (right) performance over different blur kernels. Scores are computed using YORK-120-ALL data set. All saliency maps are resized to [511 681] before blurring.

averaging kernels. As shown in Fig. 16, our evaluation metrics (**SL-AUC** & **SL-KL**) is very robust to the blur effect. The difference between our best and worst **SL-AUC** performance is 0.0022 (0.31% improvement), which is relatively small and basically has no effects to the ranking order of the model. The results also indicate that only a proper blurring operation can improve the performance, over-blurring will probably lead to performance degeneration of the model.

D. Effect of Scale

As our method is implemented using a patch-based image representation, the effect of scale (the size of patch) becomes an interesting and critical issue in our framework. In Table IV, we compared the **SL-AUC** and **SL-KL** performance of 4 single scales with the multi-scale approach on YORK-120-ALL Dataset. Our Multi-scale method (Ours-MS) outperformed all other schemes in 35% of the testing cases. The second best scheme was $[3\times3]$ which outperforms others in 21.67% cases. $[1\times1]$, $[5\times5]$ and $[7\times7]$ covered 14.17%, 14.17% and 15% of the rest, which were almost evenly distributed. Fig. 17 shows example images that are favored by different scales, which can help us better understand the scale effect to our model. In each column, we present 2 images on which the corresponding scale scheme outperforms all the others. A red box is presented in each column which indicates the size of patches sampled under the corresponding scheme. The results indicate that our current multi-scale fusion scheme indeed outputs overall better results, yet it still can not beat the best single scale scheme in 65% of the cases, because the salient patterns in these cases are mostly gathered in a certain scale.

TABLE IV
EVALUATION RESULTS OVER SCALE

	$[1 \times 1]$	$[3 \times 3]$	$[5 \times 5]$	$[7 \times 7]$	Ours-MS
SL-AUC	0.6684	0.7112	0.7100	0.6815	0.7153
SL-KL	0.1792	0.2978	0.2980	0.2259	0.3141
IL-AUC	0.6690	0.7110	0.7104	0.6837	0.7166
IL-KL	1.0649	1.3657	1.4380	1.2388	1.5186

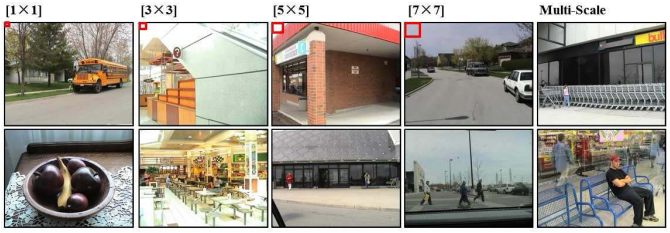

Fig. 17. Images favored by different scales. In each column, we present 2 images on which the corresponding scale scheme outperforms all the others. We also show a red box in each column which indicates the size of patches sampled under the corresponding scheme.

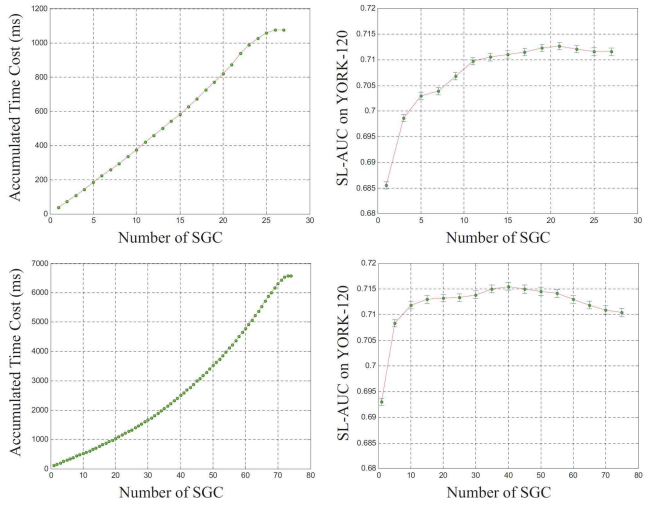

Fig. 18. The accumulated time cost (left) and SL-AUC performance (right) over different number of super Gaussian components. The testing results are obtained under 2 fixed scales including $[3 \times 3]$ (top) and $[5 \times 5]$ (bottom).

E. Effect of SGC Number

Decomposing less SGCs can significantly reduce the time cost but might also cause performance degeneration. How many SGCs should be decomposed is a specific question raised by our model. Fig. 18 shows how the time cost and the performance change over different SGC numbers. Our results indicate that it is not always the more the better. For $[5 \times 5]$, the best SGC number is 40, which achieved 3.25% SL-AUC improvement compared to the case using one single SGC. Similarly, for $[3 \times 3]$, the best SGC number is 21 which achieves 3.95% SL-AUC improvement. The single SGC scheme can reduce about 95% time cost (extreme speed = 33fps) with little performance degeneration, which is a big advantage from technical point of view.

VII. CONCLUSIONS AND FUTURE WORKS

In this paper, we introduce a unified statistical model for saccadic eye movements and visual saliency. Different from previous works that mostly aim to reproduce the exact mechanisms of visual perception, we draw inspirations from the statistical characteristics of real-world human behavior. Experimental results demonstrate our superior performance over the state-of-the-art approaches and implies the promising potential of statistical models for human behavior analysis. We also demonstrate the plausibility and effectiveness of the proposed model as well as our evaluation schemes by conducting extensive investigation on 5 key issues of saliency modeling research including evaluation scheme, model effectiveness, the effect of blur, scale and number of features.

In further studies, we will continue our effort to analyze human saccadic behavior based on statistical techniques. Our recent works indicate that, in top-down scenarios, most of the super Gaussian components will become outliers since they are irrelevant to the given goals or tasks. Thus, the statistical model for top-down attention will require new priors and new pursuit strategies which could help identify and extract the true task-relevant **SGC**s. Applying the proposed framework to other computer vision problems such as anomaly detection and pattern discovery *etc*. will also be an important direction of our future works.

REFERENCES

[1] 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.
[2] 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.
[3] D. Gao, S. Han, and N. Vasconcelos, "Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 6, pp. 989–1005, Jun. 2009.
[4] X. Sun, H. Yao, and R. Ji, "What are we looking for: Towards statistical modeling of saccadic eye movements and visual saliency," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 1552–1559.
[5] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiol.*, vol. 4, no. 4, pp. 219–227, 1985.
[6] A. Hyvärinen, J. Karhunen, and E. Oja, *Independent Component Analysis*, vol. 46. New York, NY, USA: Wiley, 2004.
[7] A. Hyvärinen, J. Hurri, and P. Hoyer, "Natural image statistics," in *Natural Image Statistics: A Probabilistic Approach to Early Computational Vision, Computational Imaging and Vision*, vol. 39. London, U.K.: Springer-Verlag, 2009.
[8] J. K. Tsotsos, S. M. Culhane, W. Y. K. Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artif. Intell.*, vol. 78, no. 1, pp. 507–545, 1995.
[9] 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.
[10] L. Itti and C. Koch, "Computational modelling of visual attention," *Nature Rev. Neurosci.*, vol. 2, no. 3, pp. 194–203, 2001.
[11] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, pp. 155–162.
[12] X. Hou and L. Zhang, "Dynamic visual attention: Searching for coding length increments," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, Inc., 2009, pp. 681–688.
[13] W. Wang, Y. Wang, Q. Huang, and W. Gao, "Measuring visual saliency by site entropy rate," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2010, pp. 2368–2375.
[14] W. Wang, C. Chen, Y. Wang, T. Jiang, F. Fang, and Y. Yao, "Simulating human saccadic scanpaths on natural images," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 441–448.
[15] L. Itti and P. F. Baldi, "Bayesian surprise attracts human attention," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2006, pp. 547–554.

[16] D. Gao, V. Mahadevan, and N. Vasconcelos, "The discriminant center-surround hypothesis for bottom-up saliency," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Associates, Inc., 2007, pp. 1–8.
[17] L. Duan, C. Wu, J. Miao, L. Qing, and Y. Fu, "Visual saliency detection by spatially weighted dissimilarity," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 473–480.
[18] 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.
[19] N. Riche, M. Mancas, M. Duvinage, B. Gosselin, and T. Dutoit, "RARE2012: A multi-scale rarity-based saliency detection with its comparative statistical analysis," *Signal Process., Image Commun.*, vol. 28, no. 6, pp. 642–658, 2013.
[20] A. Borji and L. Itti, "Exploiting local and global patch rarities for saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2012, pp. 478–485.
[21] A. Yarbus, *Eye movements and Vision*. New York, NY, USA: Plenum, 1967.
[22] M. Hayhoe, A. Shrivastava, R. Mruczek, and J. Pelz, "Visual memory and motor planning in a natural task," *J. Vis.*, vol. 3, no. 1, p. 6, 2003.
[23] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *Trends Cognit. Sci.*, vol. 9, no. 4, pp. 188–194, 2005.
[24] M. Land, N. Mennie, and J. Rusted, "The roles of vision and eye movements in the control of activities of daily living," *Perception-London-*, vol. 28, no. 11, pp. 1311–1328, 1999.
[25] L. Renninger, P. Verghese, and J. Coughlan, "Where to look next? Eye movements reduce local uncertainty," *J. Vis.*, vol. 7, no. 3, p. 6, 2007.
[26] T. Foulsham and G. Underwood, "What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition," *J. Vis.*, vol. 8, no. 2, p. 6, 2008.
[27] B. Tatler, M. Hayhoe, M. Land, and D. Ballard, "Eye guidance in natural vision: Reinterpreting salience," *J. Vis.*, vol. 11, no. 5, p. 5, 2011.
[28] T. Judd, K. Ehinger, F. Durand, and A. Torralba, "Learning to predict where humans look," in *Proc. IEEE 12th Int. Conf. Comput. Vis. (ICCV)*, Oct. 2009, pp. 2106–2113.
[29] B. Tatler, R. Baddeley, and I. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vis. Res.*, vol. 45, no. 5, pp. 643–659, 2005.
[30] A. M. Treisman and G. Garry, "A feature-integration theory of attention," *Cognit. Psychol.*, vol. 12, no. 1, pp. 97–136, 1980.
[31] P. Huber, "Projection pursuit," *Ann. Statist.*, vol. 13, no. 2, pp. 435–475, 1985.
[32] A. Hyvärinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Compupt.*, vol. 9, no. 7, pp. 1483–1492, 1997.
[33] N. Bruce and J. Tsotsos, "Saliency, attention, and visual search: An information theoretic approach," *J. Vis.*, vol. 9, no. 3, pp. 1–24, 2009.
[34] C. Guo, Q. Ma, and L. Zhang, "Spatio-temporal Saliency detection using phase spectrum of quaternion Fourier transform," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2008, pp. 1–8.
[35] M. Rubinstein, A. Shamir, and S. Avidan, "Improved seam carving for video retargeting," *ACM Trans. Graph.*, vol. 27, no. 3, p. 16, 2008.
[36] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell, "Sun: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, pp. 1–20, 2008.
[37] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2007, pp. 1–8.
[38] N. Murray, M. Vanrell, X. Otazu, and C. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 433–440.
[39] N. Murray, M. Vanrell, X. Otazu, and C. Parraga, "Low-level spatio-chromatic grouping for saliency estimation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 11, pp. 2810–2816, Nov. 2013.
[40] H. Seo and P. Milanfar, "Nonparametric bottom-up saliency detection by self-resemblance," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops (CVPR)*, Jun. 2009, pp. 45–52.
[41] A. Torralba, M. S. Castelhano, A. Oliva, 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.
[42] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1597–1604.
[43] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Advances in Neural Information Processing Systems*. Cambridge, MA, USA: MIT Press, 2006, pp. 545–552.
[44] J. Li, M. D. Levine, X. An, X. Xu, and H. He, "Visual saliency based on scale-space analysis in the frequency domain," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 4, pp. 996–1010, Apr. 2013.
[45] 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.
[46] B. Schauerte and R. Stiefelhagen, "Quaternion-based spectral saliency detection for eye fixation prediction," in *Proc. 12th Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 116–129.
[47] R. A. Rensink, "The dynamic representation of scenes," *Vis. Cognit.*, vol. 7, nos. 1–3, pp. 17–42, 2000.
[48] K. P. Balanda and H. MacGillivray, "Kurtosis: A critical review," *Amer. Statist.*, vol. 42, no. 2, pp. 111–119, 1988.
[49] R. B. Darlington, "Is kurtosis really 'peakedness?'," *Amer. Statist.*, vol. 24, no. 2, pp. 19–22, 1970.

Xiaoshuai Sun received the B.S. degree in computer science from Harbin Engineering University, Harbin, China, in 2007. He is currently pursuing the Ph.D. degree with the School of Computer Science and Technology, Harbin Institute of Technology, Harbin. He was a Research Intern with Microsoft Research Asia, Beijing, China, from 2012 to 2013, and also a recipient of the Microsoft Research Asia Fellowship in 2011. He holds two authorized patents and has authored over 40 referred journal and conference papers in the field of multimedia and computer vision.

Hongxun Yao (M'–) received the B.S. and M.S. degrees from the Harbin Shipbuilding Engineering Institute, Harbin, China, in 1987 and 1990, respectively, and the Ph.D. degree from the Harbin Institute of Technology, Harbin, in 2003, all in computer science, where she is currently a Professor with the School of Computer Science and Technology. Her research interests include pattern recognition, multimedia processing, and digital watermarking. She has published five books and over 200 scientific papers.

Rongrong Ji (M'10–SM'–) received the Ph.D. degree in computer science from the Harbin Institute of Technology, Harbin, China. He has been a Post-Doctoral Research Fellow with the Department of Electrical Engineering, Columbia University, New York, NY, USA, since 2011. He was a Visiting Student with Peking University, Beijing, China, in 2010, and a Research Intern with Microsoft Research Asia, Beijing, from 2007 to 2008. His current research interests include image and video search, and content understanding.

Xian-Ming Liu is currently pursuing the Ph.D. degree with the School of Computer Science, Harbin Institute of Technology, Harbin, China. He received the bachelor's and master's degrees in 2008 and 2010, respectively. He was a Research Intern with the Media Analyzing Group, NEC Laboratory, Beijing, China, from 2010 to 2011, and with the Web Search and Mining Group, Microsoft Research Asia, Beijing, from 2011 to 2012. His research interests include multimedia information retrieval, image/video annotation, machine learning, and computer vision.