Journal of the Optical Society of America A
OPTICS, IMAGE SCIENCE, AND VISION

Cauchy graph embedding based diffusion model for salient object detection

YIHUA TAN,[1,*] YANSHENG LI,[1] CHEN CHEN,[2] JIN-GANG YU,[1] AND JINWEN TIAN[1]

[1]National Key Laboratory of Science & Technology on Multi-spectral Information Processing, School of Automation, Huazhong University of Science and Technology, Wuhan 430074, China
[2]Department of Electrical and Computer Engineering, University of Illinois at Urbana–Champaign, Champaign, Illinois 61820, USA
*Corresponding author: yhtan@hust.edu.cn

Received 18 January 2016; revised 10 March 2016; accepted 15 March 2016; posted 15 March 2016 (Doc. ID 255593); published 18 April 2016

Salient object detection has been a rather hot research topic recently, due to its potential applications in image compression, scene classification, image registration, and so forth. The overwhelming majority of existing computational models are designed based on computer vision techniques by using lots of image cues and priors. Actually, salient object detection is derived from the biological perceptual mechanism, and biological evidence shows that the spread of the spatial attention generates the object attention. Inspired by this, we attempt to utilize the emerging spread mechanism of object attention to construct a new computational model. A novel Cauchy graph embedding based diffusion (CGED) model is proposed to fulfill the spread process. Combining the diffusion model and attention prediction model, a salient object detection approach is presented through perceptually grouping the multiscale diffused attention maps. The effectiveness of the proposed approach is validated on the salient object dataset. The experimental results show that the CGED process can obviously improve the performance of salient object detection compared with the input spatial attention map, and the proposed approach can achieve performance comparable to that of state-of-the-art approaches.

OCIS codes: (100.2960) Image analysis; (100.5010) Pattern recognition; (150.1135) Algorithms.

http://dx.doi.org/10.1364/JOSAA.33.000887

1. INTRODUCTION

A vision system always selects an important subset of visual information for further processing according to visual saliency, which is called selective attention. Whether selective attention should be spatial-based or object-based is debated in the field of visual attention [1–3]. Correspondingly, the saliency computation models in the visual saliency field can be classified into two categories: one designed for fixation prediction [4,5], the other one designed for salient object detection [6,7]. The approaches in the former category, inspired by biological vision, are used to predict the spatial location where the attention of human eyes will focus in a scene. On the other hand, the latter category is devised based on techniques of computer vision for salient object detection in the complex scenes. Because the detection of object from images has extensive applications, the detection framework following visual attention has been one of the mainstream strategies in recent years.

Most saliency computation models for fixation prediction were motivated by simulating the eye motion of biological vision in order to extract the salient point for attracting attention [4,5]. Inspired by physiological conclusions of the primary visual cortex, Itti *et al.* proposed to create the final saliency map by linearly combining the differential results which were output from multichannel and multiscale data by a center–surround operator [8]. Ma and Zhang presented a local contrast saliency model based on fuzzy growth [9]. Harel *et al.* defined the saliency with a normalized center–surround differential image in order to emphasize the salient regions [10]. Zhang and Sclaroff decomposed the original image into a group of bivalue images where the attention results were computed, and then combined the attention maps into the final saliency map [11]. Even though the performance of fixation prediction models has been much improved in recent years, these approaches are hard to apply to object detection tasks directly as the predicted focus generally locates on the edges or corners.

On the other hand, inspired by object-based attention, salient object detection concentrates on the prediction and segmentation of an object in an image [6,7]. Achanta *et al.* provided a salient object prediction approach based on frequency modulation and constructed a database for the evaluation of salient object detection approaches [12]. Cheng *et al.* devised a salient object prediction algorithm using global region contrast, in which the salient objects were segmented with an iterative graphical model [13]. Perazzi *et al.* presented a filtering algorithm for salient object detection by integrating local and global contrast [14]. Observing that a salient object is not

totally distributed in the boundary area, the researchers exploited the knowledge of boundaries in saliency detection [15,16]. Recently, Zhu *et al.* proposed a graph optimization model to fuse multiple salient clues so as to obtain a more intact object [17]. Generally speaking, the main purpose of such salient object detection approaches is to perceptually group the different parts of an object as a smooth region distinct from the background. However, the saliency measures are normally computed according to the contrast of pixel strength, which makes it difficult to segment the salient object in an image because of the saliency discrepancy among different object parts.

In order to increase the smoothness of the saliency map for object detection, a lot of researches in the literature have been conducted to implement the saliency diffusion or propagation. Actually, the interpretation about the mechanism of object-based attention also provides diffusion process a reasonable biological foundation. The mechanism of sensory enhancement emphasizes that the attentional object is selected by spreading the attention until the boundary of the attended object according to Gestalt principle [3]. The sensory enhancement theory was deeply investigated [18,19], and the neurophysiological evidences have been be found in [20]. The attention spread process, which is the result of modulating object-based attention according to the sensory enhancement theory, is able to be modeled as the diffusion process of propagating the initial states (spatial attention) from some nodes to the other nodes along a connected graph. In [21], the state propagation from a node to another node can be explained from the angle of PDE-based anisotropic diffusion. Incorporating human perception and high-level human prior, Liu *et al.* proposed an approach that the saliency evolution from the salient seeds to the relevant points is modeled as an optimal partial differential equation, which can be learned from an image [22]. Jiang *et al.* formulated saliency detection via an absorbing Markov chain on an image graph model, where the absorbed time of the transient node is used to measure the saliency [23]. Ren *et al.* combined a compactness metric and modified PageRank propagation to represent the object saliency in a perceptually meaningful way [24]. Gong *et al.* believed that the optimization of the saliency propagation should deal with the inhomogeneous difficult adjacent regions [25]. Therefore, they employed the teaching-to-learn and learning-to-teach strategies to improve the propagation quality.

From the viewpoint of graph embedding, the saliency of a node can be considered as the low-dimension embedding of the high-dimension appearance feature, which can be composed of Gestalt cues as explained in [3]. Specifically, the nodes represented by features are embedded into one-dimensional (1D) space with states as coordinates [26]. Combining the initial saliency map with the embedding process, we can build a novel diffusion model for the saliency propagation. The key step for this model is to choose the best graph embedding to preserve the local topology in both the original and the embedding space. Therefore, we simply review the techniques about graph embedding before we build our diffusion model.

As a kind of unsupervised dimension reduction method, graph embedding can be classified into two categories: linear embedding approaches [27,28] and nonlinear embedding approaches [29–31]. The first type of approaches conduct the low-dimension embedding of data through linear transformation, including principal component analysis [27] and multidimensional scaling [28], which were extensively applied in computer vision and machine learning fields. However, the linear embedding cannot achieve good performance when the actual data has inherent nonlinear structures. To achieve better performance, nonlinear embedding approaches were proposed, such as IsoMap (preserving the global topological structure) [29], local linear embedding (preserving local topological structure) [30], and Laplacian embedding [31]. Among the nonlinear embedding models, quadratic functions are adopted to define the loss [32–34]. Laplacian embedding was regarded as one of the most successful nonlinear embedding approaches [31]. Luo and co-workers proposed the Cauchy graph embedding to better preserve the local topology [26,35]. As the local topology is very important for the state propagation in the attention spread process, we construct the diffusion model by using Cauchy graph embedding.

Inspired by the sensory enhancement theory [3], not to simulate the mechanism, this paper presents a salient object detection framework. As the foundation of our detection framework, we propose a Cauchy graph embedding based diffusion (CGED) model to conduct saliency propagation. To cope with the propagation between the inhomogeneous regions, we instantiate CGED as a fitting diffusion model and merging diffusion model, which have different propagation ability. Integrated with the diffusion model, the framework includes the following steps: (1) oversegment the original image on multiple scales; (2) produce attention prediction maps according to the multiscale oversegmented images; (3) diffuse the attention maps using fitting diffusion and merging diffusion models consequently; (4) perceptually group the multiscale diffused salient maps.

In general, the contributions of the paper can be summarized as follows: (1) propose a diffusion model by combining the initial saliency state and graph embedding; (2) realize the saliency spread by applying the diffusion model based on Cauchy graph embedding; (3) design a two-stage diffusion strategy to cope with the propagation between inhomogeneous regions; (4) implement the diffusion process by an iterative gradient descent algorithm. The experimental results show the advantages of our algorithm: (1) the Cauchy embedding based diffusion model is more appropriate for the attention spread compared with the Laplacian embedding based diffusion model, and (2) the performance of salient object prediction by the visual grouping of Cauchy graph embedding can significantly improve the original spatial attention result and is comparable to state-of-the-art algorithms.

The remaining parts of the paper are organized as follows. Section 2 explains the mathematical formula and optimization solution of the diffusion model based on graph embedding. Section 3 introduces the salient object detection algorithm by utilizing the CGED model. In Section 4, we first provide the evaluation dataset and evaluation principles, then analyze the rationale of the proposed CGED model, and finally give the comparative results of our algorithm and state-of-the-art salient object detection approaches. Section 5 is the conclusion of the paper.

2. CGED MODEL

Inspired by the sensory enhancement theory that the attention spread mechanism is the main argument of object-based attention, we attempt to implement the process using the diffusion model based on graph embedding. The diffusion process is equivalent to the data embedding from high-dimension feature space to 1D state space while we preserve the initial states as far as possible. Thus, the optimization equation consists of two terms: a fitting term and a regularized smooth term.

To facilitate the illustration of the graph embedding based diffusion model, the construction of the graph model is first introduced. Let $\mathbf{G} = (\mathbf{V}, \mathbf{E})$ represent the undirected graph, where $\mathbf{V}$ is the node set and $\mathbf{E}$ is the edge set denoting the connection relation. The edge weight $w_{i,j}$ between node i and j can be expressed by

$$w_{i,j} = \exp\left(-\frac{\|\|h_i - h_j\|\|^2}{2\sigma_w^2}\right), \tag{1}$$

where h_i and h_j are the high-dimensional descriptive features of nodes i and j, respectively, and σ_w is the weight control constant. The descriptive feature can also be viewed as the Gestalt cues that we can extract from the original image.

In order to maintain the same local structure both in the high-dimension feature and the 1D state spaces, the property of preserving local topology is very crucial. Therefore, it is critical to select the appropriate graph embedding to conduct the diffusion process. In the following sections, we give the details of the diffusion model based on graph embedding and the optimization algorithm.

A. Laplacian Graph Embedding Based Diffusion Model

Suppose the given graph has n nodes; $\mathbf{y} = [y_1, y_2, \ldots, y_n]^T$ denotes the initial states of n nodes. The initial state vector $\mathbf{y}$ is generally contaminated by noise without considering the inherent geometric topological structure. It is natural to restore the genuine state of each node by which the initial state of a node spreads along the inherent geometrical structure. Within the regularization framework of [36], one way to simulate the spread is that we can use Laplacian embedding [31] as the regularized smooth term to approximate the diffusion effect. Laplacian graph embedding based diffusion (LGED) is formulated as Eq. (2), and the states of n nodes $\mathbf{f}^* = [f_1^*, f_2^*, \ldots, f_n^*]$ come into equilibrium when the energy of Eq. (2) is minimized. Here, we choose the quadratic function to measure the loss. In Eq. (2), we drop out the additional constraint in Laplacian eigenmaps [31] because the initial state $\mathbf{y}$ can be viewed as the anchor point which can avoid the arbitrary scale factor in the embedding of f^*. Thus the LGED is expressed as

$$\mathbf{E}(\mathbf{f}) = \sum_i (f_i - y_i)^2 + \lambda \sum_{i,j} w_{i,j}(f_i - f_j)^2, \tag{2}$$

where the first fitting term encourages the node to maintain its initial state, and the second smooth term as embedding expression forces the nodes to take the similar states for those similar nodes in high-dimension space. λ is the punishment factor. The small λ makes the optimized state tend to be close to the initial value, and the big λ leads to the final state being in accordance with the inherent topological structure of the data.

By setting the derivative of Eq. (2) corresponding to $\mathbf{f}$ as zero, the final state of equilibrium is expressed as

$$\mathbf{f}^* = u(\mathbf{D} - \mathbf{W} + u\mathbf{I})^{-1}\mathbf{y}, \tag{3}$$

where $u = 1/2\lambda$, $\mathbf{D} = \mathrm{diag}(d_1, d_2, \ldots, d_n)$, $\mathbf{W}$ is the weight matrix constructed by $w_{i,j}$, $\mathbf{I}$ is the unit matrix, and $d_i = \Sigma_j w_{i,j}$.

If we replace the regularization term of Eq. (2) with manifold ranking, which can preserve the local and global consistency [32,37], the manifold ranking based diffusion (MRD) model can be defined as

$$\mathbf{E}(\mathbf{f}) = \sum_i (f_i - y_i)^2 + \lambda \sum_{i,j} w_{i,j}\left(f_i/\sqrt{d_i} - f_j/\sqrt{d_j}\right)^2. \tag{4}$$

Similar to the optimization of Eq. (2), the final state of equilibrium can be formulated as

$$\mathbf{f}^* = (\mathbf{I} - u\mathbf{S})^{-1}\mathbf{y}, \tag{5}$$

where $u = \lambda/(\lambda + 1)$, $\mathbf{S} = \mathbf{D}^{-1/2}\mathbf{W}\mathbf{D}^{-1/2}$ denotes a normalized Laplacian matrix.

B. CGED Model

As shown in LGED [as formula (2)] and MRD [as formula (4)], the fitting term and smooth term take a quadratic function to measure the loss. However, a quadratic loss function is disadvantageous to preserve the local topology of data [26,35], which leads us to introduce a bounded function to define the loss. The diffusion model using bounded function is defined as

$$\mathbf{E}(\mathbf{f}) = \sum_i \frac{(f_i - y_i)^2}{(f_i - y_i)^2 + \delta^2} + \lambda \sum_{i,j} w_{i,j} \frac{(f_i - f_j)^2}{(f_i - f_j)^2 + \delta^2}, \tag{6}$$

where δ is a sensitivity control factor, and the determination of its value will be discussed in Section 3. Through eliminating $(f_i - y_i)^2$ and $(f_i - f_i)^2$ in the numerator by adding δ^2 and subtracting δ^2, the minimization of Eq. (6) is equivalent to the maximization of Eq. (7):

$$\mathbf{E}(\mathbf{f}) = \sum_i \frac{1}{(f_i - y_i)^2 + \delta^2} + \lambda \sum_{i,j} w_{i,j} \frac{1}{(f_i - f_j)^2 + \delta^2}. \tag{7}$$

Since the Cauchy distribution function $f(x) = 1/(x^2 + \delta^2)$ is adopted in formula (7), which is a simple variant of Eq. (6), the model defined using Eq. (6) is named CGED. Compared with the quadratic loss function $f(x) = x^2$ used in Eq. (2), the bounded function $f(x) = x^2/(x^2 + \delta^2)$ is able to avoid the small amount of the highly dissimilar points playing the leading role during the data embedding. Figure 1 shows that $f(x) = x^2$ is an unbounded and monotonically increasing function; meanwhile, $f(x) = x^2/(x^2 + \delta^2)$ is also monotonic but bounded. The quadratic function as loss function puts emphasis on the property of partitioning the long distance point pairs but ignores the property of preserving the topology of small distance point pairs. As a result, a large amount of small distance point pairs are unable to contribute to the preservation of the local topological structure. In contrast, by

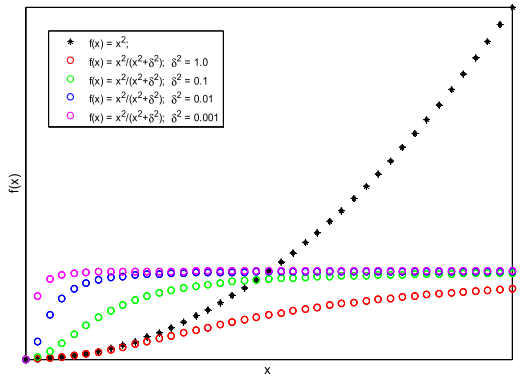

Fig. 1. Variation trend of different loss functions.

bounding the loss function, the bounded function can avoid distorting the topology of data by limiting the finite contribution of the highly dissimilar point pairs. The example of the local topology preserving property was demonstrated in [26]. In Fig. 2, we show that the bounded function as a loss function can encourage the similar superpixels in the color space to be close in the embedding space. As can be seen from Fig. 2, the values of two similar superpixels pairs embedded with the bounded function are closer than that of the corresponding pair embedded with the quadratic function. This property is very crucial for salient object detection because it can make the inside of the object homogeneous.

Because the solution to the problem of Eq. (7) has no close form, the next section will demonstrate the details of the optimization steps based on a gradient descent algorithm.

C. Optimization of CGED Model

In this section we obtain the state of equilibrium $\mathbf{f}^*$ by maximizing the energy function $\mathbf{E(f)}$ of Eq. (7) using a gradient descent algorithm. Equation (8) denotes the gradient of energy function $\mathbf{E(f)}$ corresponding to the optimization variable $\mathbf{f}$:

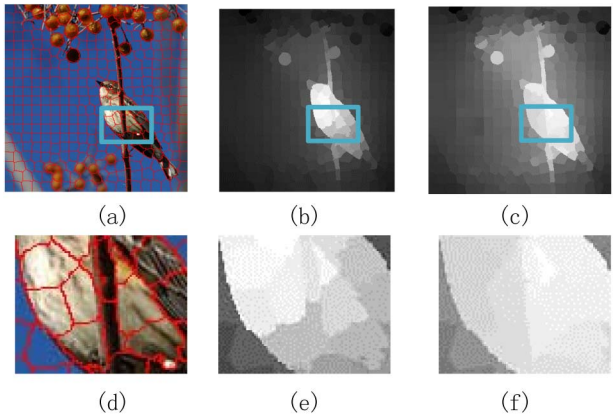

Fig. 2. Embedded results for saliency computation. (a) Superpixel segmentation result; (b) embedded result using quadratic function; (c) embedded result using bounded function. (d), (e), and (f) are the enlarged version of the local rectangle area corresponding to (a), (b), and (c), respectively.

$$\frac{\partial \mathbf{E(f)}}{\partial f_i} = -2\delta^2 \frac{(f_i - y_i)}{((f_i - y_i)^2 + \delta^2)^2} - 2\lambda\delta^2 \sum_j w_{i,j} \frac{(f_i - f_j)}{(f_i - f_j)^2 + \delta^2}. \tag{8}$$

Therefore, the specific optimization steps of the CGED model is detailed as Algorithm 1. The addition of momentum term in gradient descent is one approach for decreasing the rate of convergence dramatically [38]. This approach is also very popular in deep learning algorithms. Since the momentum is normally between 0.5 and 0.9, the decreasing amounts of those time steps before $t-1$ have comparative smaller contribution to the optimal f^*. For simplicity, in our implementation we only consider the decreasing amounts of f at time step $t-1$ and t in the updated Eq. (9).

Algorithm 1 Optimization for CGED model

Initialization: $\mathbf{y}$ is the input initial state, t is the iteration index, let $\frac{\partial \mathbf{E}^{t-1}(\mathbf{f})}{\partial f_i} = \frac{\partial \mathbf{E}^{t}(\mathbf{f})}{\partial f_i} = 0$, $t = 2$, $i = 1, 2, \ldots, n$, set the gradient descent step ε, momentum v, and $\mathbf{f} = \mathbf{y}$.

Repeat

1) Compute $\frac{\partial \mathbf{E}^{t}(\mathbf{f})}{\partial f_i}$ according to Eq. (8), $i = 1, 2, \ldots, n$, and then normalize them.
2) Update:

$$f_i^{t+1} = f_i^t - \varepsilon\left((1-v)\frac{\partial \mathbf{E}^{t-1}(\mathbf{f})}{\partial f_i} + v\frac{\partial \mathbf{E}^{t}(\mathbf{f})}{\partial f_i}\right), \qquad i = 1, \ldots, n \tag{9}$$

Until $\mathbf{E(f)}$ converges.

Output: the optimization state of equilibrium $\mathbf{f}^* = \mathbf{f}^{t+1}$.

3. SALIENT OBJECT DETECTION USING CGED MODEL

In this section, we present a novel salient object detection framework using the diffusion model introduced in the preceding section. It comprises four subsections: the framework of salient object detection, Cauchy graph embedding based fitting diffusion (CGEFD), Cauchy graph embedding based merging diffusion (CGEMD), and multiscale perception grouping.

A. Framework of Salient Object Detection

Inspired by sensory enhancement theory, but not to simulate the mechanism precisely, the proposed algorithm focuses on the spread of spatial attention from the initial state. In order to avoid producing the spatial attention result by hand, classical human attention focus prediction is regarded as the spatial attention which is also the initial input of our presented detection algorithm. The flowchart of the presented algorithm is shown as Fig. 3: The original image is first multiscale segmented, and spatial attention is deduced according to the multiscale oversegmented image, then fitting diffusion and merging diffusion are conducted on the initial attention maps sequentially, and finally the multiscale diffusion results are combined to implement perceptual grouping of salient object. The multiscale strategy in the computer vision field makes the algorithm more robust.

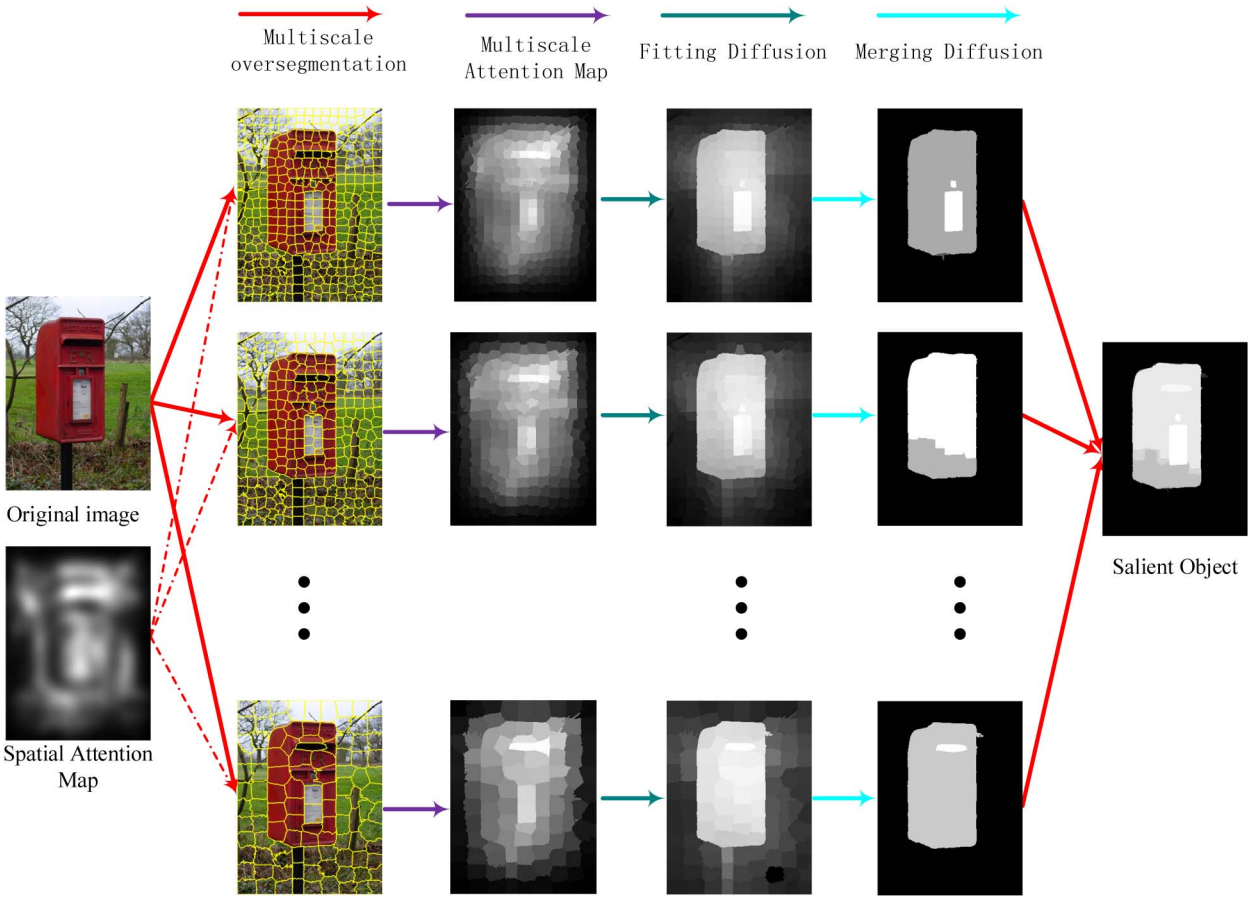

Fig. 3. Flowchart of salient object detection using CGED model, where the spatial map is produced automatically using the human eye fixation prediction approach proposed by Itti *et al.* [8].

For the first step in this detection framework, the simple linear iterative clustering (SLIC) algorithm proposed by Achanta *et al.* [39] is adopted to produce near uniform size superpixels as the elementary computation units for the attention spread. In the SLIC algorithm there are two important parameters: compactness c and number of segmentation blocks n. The multiscale perception is realized by changing the block number n. In the experimental setting, compactness c is set to 25 for the sake of boundary precision of superpixels, and n is set to 100, 200, 300, and 400 in order to generate superpixel representation with four scales.

For the second step of the detection framework, two classical human eye focus prediction models, including the biologically inspired prediction model proposed by Itti *et al.* [8] and the graph based visual saliency (GBVS) model proposed by Harel *et al.* [10], are introduced in this paper. The approach proposed by Itti *et al.* [8] consists of two steps: visual stimulus (feature) computation using difference of Gaussian filters (DoG) on multiple channels and multiple scales, and the combination of visual features across channels and scales. Meanwhile, the approach proposed by Harel *et al.* [10] can integrate the visual feature extraction and normalization into a graph model. In Fig. 3 we take Itti's model as an example to explain the flowchart of our proposed algorithm. In the experimental section, we also give the salient object detection results when the initial attention map is produced using the GBVS model in order to prove the robustness of our algorithm.

For the third step of the framework, as the core part to fulfil the spread task of spatial attention, we choose the CGED model to implement the diffusion process from the initial spatial attention map. As pointed out in Section 2.B, different δ also has a distinct impact on the final property of preserving topology. Therefore, we use two diffusion stages to conduct the attention spread process. The first one is named the fitting diffusion model, which prefers to process the superpixels with big dissimilarity, when δ^2 is set to 1. This step can decide the basic boundary of an object. The second one is named the merging diffusion model, which inspires the superpixels with small dissimilarity to be merged together in the saliency map, when δ^2 is set to 0.0001. This step can smooth the inner part of an object. The details are described in Sections 3.B and 3.C.

During the diffusion process, the edge weight is obtained from the feature similarity according to Eq. (1). Since we work on the basis of superpixels, we describe the feature extraction of each superpixel as follows. Given the parameter set, the original image is segmented into n regions with near uniform size $\{r_i | i = 1, 2, \ldots, n\}$. In accordance with the biological perception, the first and second order statistics are combined to describe the features of each region [40–42]. For each pixel

(x, y) in region r_i, the descriptive features are represented as $(x, y, L, a, b, |I_x|, |I_y|, |I_{xx}|, |I_{yy}|, |I_{xy}|)$, where (L, a, b) is the value of a pixel in Lab color space, and $(|I_x|, |I_y|, |I_{xx}|, |I_{yy}|, |I_{xy}|)$ are first-order derivatives in the x and y directions, second-order derivatives in the x and y directions, and second-order partial derivatives in both the x and y directions. For each region r_i, the first-order descriptive feature is represented by the mean $\mathbf{m}_i$ of the feature vectors of all the pixels belong to r_i, and the second-order descriptive feature is denoted as the covariance $\mathbf{C}_i$ of the feature vectors in this region. For the sake of measuring the similarity of the features, the descriptive feature of a region r_i is characterized by combining the mean vector and the covariance matrix as

$$\mathbf{h}_i = [\mathbf{m}_i, s_i^1, s_i^2 \ldots, s_i^{10}], \tag{10}$$

where s_i^j is the jth vector of the Cholesky decomposition of the covariance matrix $\mathbf{C}_i$.

The final step of the detection framework is to perceptually group the multiscale saliency maps to a merged saliency map. Section 3.D gives the specific introduction of the strategy.

B. CGEFD Model

Suppose the original image is segmented into n regions with near uniform size $\{r_i | i = 1, 2, \ldots, n\}$ and the spatial locations of attention are formed through the human eye focus prediction approach, we consider each region as a node of a graph; the weight of the edge connecting two nodes is represented as the similarity between the features of two regions. Then the spatial locations of attention are mapped to the oversegmented image, and the mean of each region is evaluated as the initial state of each node. Let the initial state of n nodes be denoted as $\mathbf{y} = [y_1, y_2, \ldots, y_n]$ and the similarity matrix of nodes denoted as $\mathbf{W}$.

As shown in Fig. 1, the loss function $f(x) = x^2/(x^2 + \delta^2)$ changes gently in the region of big values when the sensitivity control factor is big. The diffusion equation corresponding to this loss function encourages the nodes to trivially diffuse (fitting diffusion) from the initial states except for those superpixels with big dissimilarity. Therefore, the CGED is instantiated as CGEFD if the sensitivity control factor is large enough. In this paper, we set δ^2 as 1.0 and put the initial state $\mathbf{y}$ and similarity matrix $\mathbf{W}$ into Algorithm 1. As a result, the diffusion map from CGEFD is generated from the equilibrium state $\mathbf{f}^*$ of n nodes. As shown in Fig. 4(b), the diffusion map brings out the contour of an object compared with the initial spatial attention. As the δ^2 decreases, more superpixels tend to be diffused accordingly, shown as Figs. 4(c) and 4(e). In Fig. 4(e) the contour of the ball has disappeared. So the best way to keep a good boundary and smooth inner part of an object is to take two steps. The second step must diffuse all the superpixels since we have already obtained the good initial contour of an object.

C. CGEMD Model

As we can see from Fig. 1, the loss function $f(x) = x^2/(x^2 + \delta^2)$ varies radically and approaches to the upper bound very quickly when the sensitivity control factor δ is small. Therefore, the diffusion equation in terms of the loss function encourages the states of all the nodes to diffuse in a merging way. In this case, CGED is instantiated as CGEMD. Taking the equilibrium state $\mathbf{f}^*$ resulting from CGEFD as the initial state, the new equilibrium state $\mathbf{f}^{**}$ of CGEMD is obtained when we set δ^2 to 0.001 and maintain the same $\mathbf{W}$. Comparing Figs. 4(e) and 4(d), the saliency map of CGEMD based on CGEFD is more compatible with the actual distribution of an object than that of CGED with $\delta^2 = 0.1$.

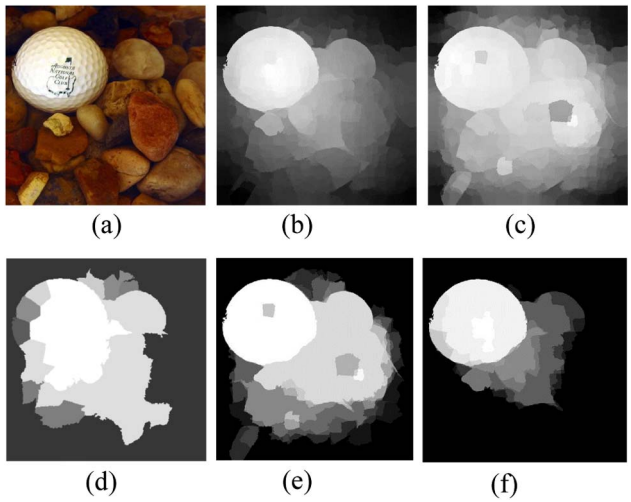

Fig. 4. Salient detection result with CGEFD and CGEMD. (a) Original image, (b) result of CGEFD ($\delta^2 = 1.0$), (c) result of CGED ($\delta^2 = 0.1$), (d) result of CGEMD ($\delta^2 = 0.001$), (e) result with two CGED steps ($\delta^2 = 0.1$ and $\delta^2 = 0.001$), (f) result with CGEFD and CGEMD ($\delta^2 = 1.0$ and $\delta^2 = 0.001$).

D. Multiscale Perception Grouping

Motivated by the success from the multiscale superpixel manipulation [43–45], this paper implements the multiscale superpixel perception grouping via the proposed CGED model.

Suppose $\mathbf{F}$ is the image representation mapped from the states $\mathbf{f}^{**}$, and $\mathbf{F}^s$ is the result of merging diffusion in scale s; then the salient object prediction map Sal according to multiscale perception grouping can be expressed as

$$\mathrm{Sal} = \frac{1}{S}(\mathbf{F}^1 + \mathbf{F}^2, + \ldots + \mathbf{F}^S), \tag{11}$$

where S is the number of scales. In our experimental settings, S is set as 4. The final salient object prediction map produced from multiscale perception grouping is shown in Fig. 2.

4. EXPERIMENTAL RESULTS

In this section, we conduct experiments on the open access evaluation dataset for salient object detection and compare the proposed algorithm with the existing state-of-the-art approaches. First, the evaluation datasets and the adopted performance measurements are introduced. Next, the experiments verify the superiority of multiscale perception grouping and the rationality of the simulation in which the CGED model is used to fulfill the mechanisms of the attention spread. Finally, the proposed algorithm is quantitatively compared with the existing state-of-the-art approaches.

In our experiments, we set the following paramters: λ is 100, σ_w is 0.1, and learning rate ε is 0.005.

A. Dataset

The dataset for performance evaluation provided by Achanta *et al.* [12] was specially designed for the salient object detection task. There are 1000 images selected from the Microsoft Research Asia (MSRA) salient object dataset [46], and the salient object in each image was labeled by hand. The ground truth dataset and original dataset can be publicly downloaded from the websites mentioned in [12] and [46], respectively.

Similar to the quantitative analysis in [12,13], a precision-recall curve is adopted to evaluate the performance of different algorithms. Specifically, after the salient map is normalized into the interval [0,255], the salient map is segmented into 256 salient object results by using the threshold incremented from 0 to 255. Therefore, the precision-recall curve can be obtained by connecting all the points that are precision-recall values through comparing the segmented objects with the ground-truth objects.

B. Advantages of CGED Model

In this paper, the human eye focus prediction models proposed by Itti *et al.* (Itti's) [8] and Harel *et al.* (GBVS) [10] are utilized to initialize the spatial attention map for our experiments. As shown in Figs. 3(c) and 3(d), the initial attention maps predicted from the models [8,10] only highlight the edges or contours of an object, which is not good enough for the extraction of an object. Basically, the core idea of our proposed algorithm is that the diffusion model is exploited to fulfil the similar function of a spatial attention spread mechanism. Through the diffusion process, the salient objects become comparatively intact in the saliency map.

To reveal the advantages of the CGED, we compare the experimental results of three models: LGED, MRD, and our proposed diffusion model (CGED). As we mentioned before, by changing the sensitivity control factor, the CGED can be instantiated to two diffusion models: CGEFD and CGEMD.

The elementary units for attention diffusion are SLIC superpixels [39] for the sake of computing efficiency. In this experiment, we only consider one scale and all the images are over-segmented into 300 superpixels. Next, the diffusion process is conducted on the superpixel based attention map which is projected from the initial spatial attention map created from human eye focus prediction models [12,14]. We have eight experimental configurations: Itti's+LGED, where the Itti's attention map is diffused using LGED, and the result is shown as Fig. 5(e); Itti's+MRD, where the Itti's attention map is diffused using MRD, and the result is shown as Fig. 5(f); Itti's+CGEFD, where the Itti's attention map is diffused using CGEFD, and the result is shown as Fig. 5(g); Itti's+CGEFD+CGEMD, where the Itti's attention map is sequentially diffused using CGEFD and CGEMD, and the result is shown as Fig. 5(h); GBVS+LGED, where the attention map of GBVS is diffused using LGED, and the result is shown as Fig. 5(i); GBVS+MRD, where the attention map of GBVS is diffused using MRD, and the result is shown as Fig. 5(j); GBVS+CGEFD, where the attention map of GBVS is diffused using CGEFD, and the result is shown as Fig. 5(k); and GBVS+CGEFD+CGEMD, where the attention map of GBVS is sequentially diffused using CGEFD and CGEMD, and the result is shown as Fig. 5(l).

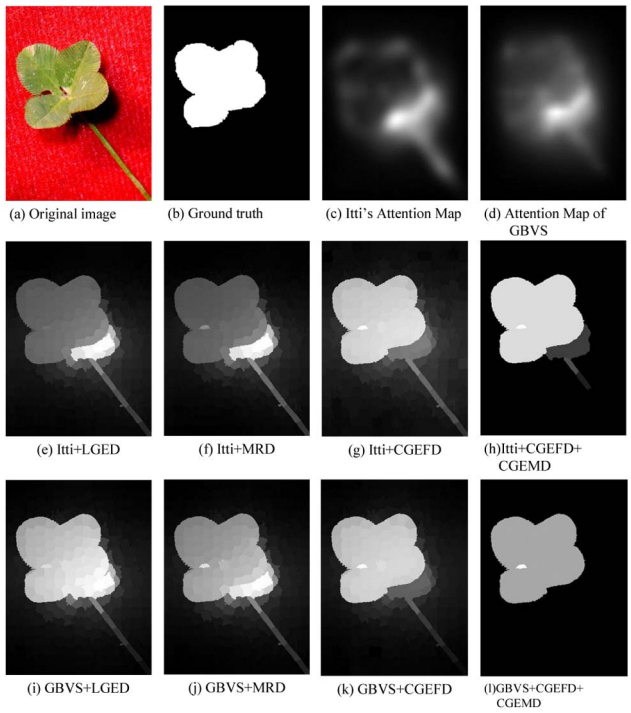

Fig. 5. Salient object prediction map using different diffusion models. (a) Original image. (b) Ground truth. (c) Itti's attention map. (d) Attention map of GBVS. (e) Itti + LGED. (f) Itti + MRD. (g) Itti + CGEFD. (h) Itti + CGEFD + CGEMD. (i) GBVS + LGED. (j) GBVS + MRD. (k) GBVS + CGEFD. (l) GBVS + CGEFD + CGEMD.

The comparison of the second and the third rows of Fig. 5 indicates that the salient object prediction performance by taking the attention map of GBVS as the initial state is better than that by adopting Itti's attention map as the initial state when we diffuse the initial map using the same model. The fact is in accordance with the evaluation performance of Itti's and GBVS models on the experimental dataset of human eye focus prediction. Therefore, we can conclude that the salient object prediction performance of our proposed algorithm may be improved if a better spatial attention prediction approach is presented in the future.

Taking the Itti's or GBVS attention map as the initial result, the salient object prediction performances of different diffusion models are quantitatively reported in Fig. 6. Generally speaking, the prediction performance is not good if only the Itti's or GBVS attention map is considered, which can be significantly improved through the diffusion based on graph embedding. As can be seen from Figs. 6(a) and 6(b), no matter what spatial attention model is taken, the performance of CGEFD is better than that of LGED and MRD, and the performance by combining CGEFD and CGEMD is better than that of CGEFD.

C. Advantages of Multiscale Perception Grouping

In this experiment, we set the superpixel number n to 100, 200, 300, and 400 to form computational units of four scales. The

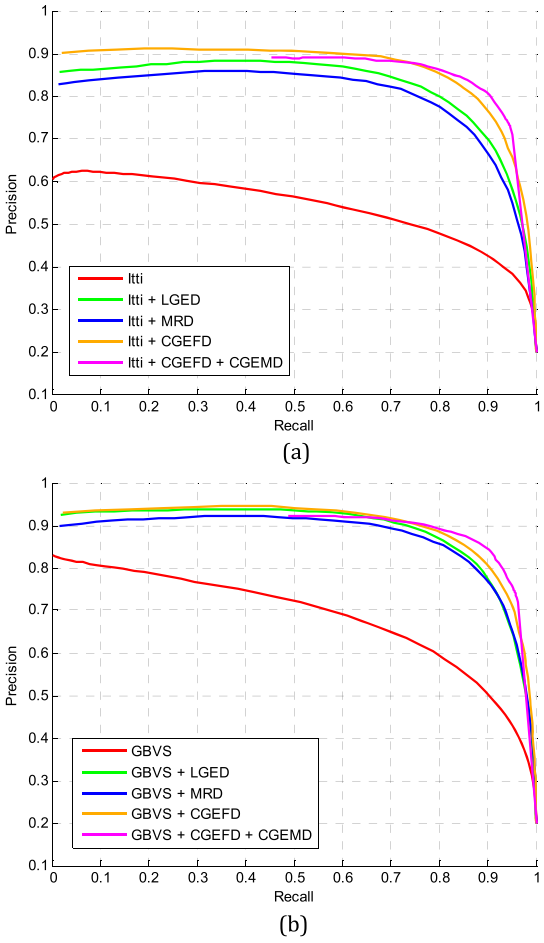

Fig. 6. Quantitative evaluation of salient object prediction based on different diffusion models. (a) Comparative results of different diffusion models when Itti's attention map is adopted as the initial map; (b) comparative results of different diffusion models when the attention map of GBVS is adopted as the initial map.

Fig. 7. Diffusion results of one scale and multiple scales. (e)–(h) One scale diffusion results on the basis of Itti's attention map; (i)–(l) multiscale diffusion results on the basis of Itti's attention map; (m)–(p) one scale diffusion results on the basis of attention map of GBVS; (q)–(t) multiscale diffusion results on the basis of attention map of GBVS.

initial attention map of each scale is diffused on the basis of superpixels. The perception grouping is conducted by fusing the diffused results of four scales. Then the fused attention map is used to predict salient object. To explain the efficiency of multiscale perception grouping, the experiment gives the visual results of one scale ($n = 300$) and multiple scales, which are shown as Fig. 7.

Furthermore, Fig. 8 quantitatively illustrates the comparative performance via one scale and multiscale perception grouping. We can conclude from the two figures that in any case, no matter which initial attention map or which diffusion model is adopted, multiscale perception grouping is more beneficial to suppress background and enhance object than one scale perception grouping in this experiment.

D. Performance Comparison with Other Approaches

In this section, we compare our proposed approach with some other approaches in the literature to verify its prediction performance. Because our proposed approach focuses on the spread of spatial attention, the approaches to be compared can be classified into two categories: spatial attention prediction models and salient object prediction models. The first category includes Itti's [8], Ma and Zhang (MZ) proposed [9], GBVS [10], and spectral residual (SR) [47], and the second category consists of 10 approaches, namely, Achanta *et al.* (AC) proposed [48], frequency-tuned (FT) saliency [12], saliency filters (SF) [14], geodesic saliency (GS) [15], graph-based manifold ranking (MR) saliency [16], saliency optimization (SO) [17], maximum symmetric surround (MSS) saliency [49], global cues (GC) saliency [50], region-based contrast (RC) saliency [51], saliency tree (ST) [45]. With the two basic configurations, multiscale GBVS+CGEFD and multiscale GBVS+CGEMFD+CGEMD, the proposed algorithm is compared with the above 13 approaches.

Among these compared algorithms, Itti's [8], AC *et al.* [48], GC [50], and ST [45] are multiscale ones.

The quantitative analysis of the above approaches with our proposed algorithm is shown in Fig. 9, and the visual comparison is illustrated in Fig. 10. Figure 9 give us the precision-recall pairs with different thresholds to segment the saliency maps. Shown as Figs. 9 and 10, the performance of our proposed algorithm is obviously more desirable than that of Itti's, MZ, GBVS, and GC prediction models, which shows that the attention spread along an object is rewarding to predict salient objects. Compared with the second category of salient object detection approaches, the proposed algorithm is superior to most of the approaches, such as AC, FT, MSS, GS, SF, and GC. The proposed algorithm also achieves comparable performance of state-of-the-art approaches, such as MR [16], RC [51], SO [17], and ST [45]. In order to explain the efficiency

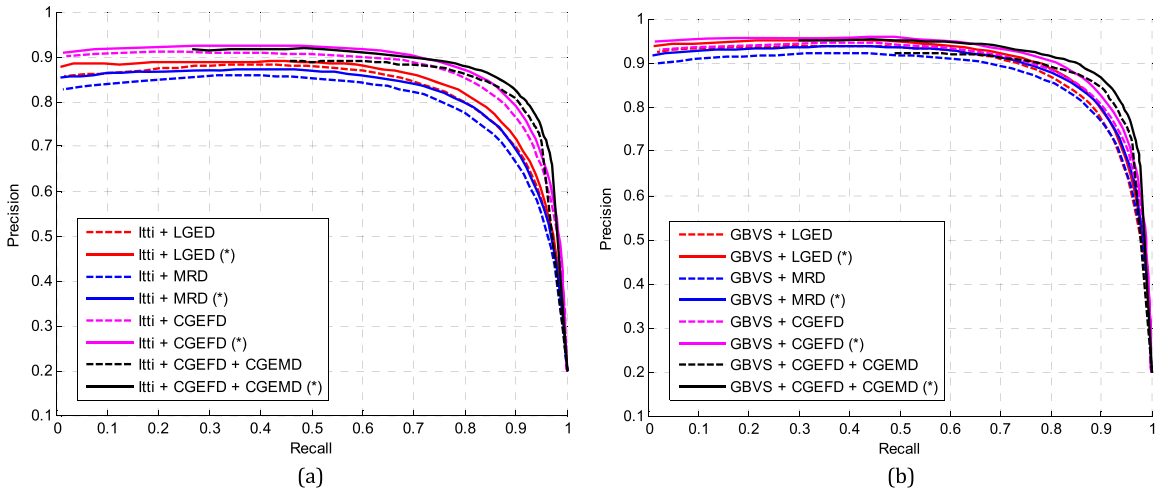

Fig. 8. Precision-recall curve of one scale and multiscale diffused results under different diffusion models. The solid line indicates the result of multiple scales and the dotted line indicates the result of one scale. (a) The initial attention map is obtained from Itti's approach; (b) the initial attention map is obtained from GBVS approach.

of salient object detection, the segmented results with the best average precision-recall on the dataset are shown as Fig. 10(s). The threshold is set as 125 when the best average recall and precision in the curve are 0.89 and 0.87, respectively.

Among these algorithms with best performance, ST and SO have almost same precision-recall curve in the interval of recall near to 0.9. Except for the multiscale strategy, the better initial grouped region with object prior maybe one of the main reasons that ST can achieve the best result among the compared algorithms.

As for our proposed algorithm, the quality of the initial attention map has also a crucial impact on the final performance of salient object prediction. Inspired by the successful strategy of ST [45], other excellent attention models with better attention prediction can be potentially integrated into our framework to further improve the performance of our method. We consider this as future work.

E. Complexity Analysis

The proposed algorithm was implemented using Matlab 2010. The superpixel segmentation used the executable program of the SLIC algorithm [39]. The Matlab source code of GBVS [10] was adopted to produce the attention map [52]. All the experiments were conducted on a laptop with Intel i7-4700HQ 2.4GHZ CPU and 4 GB RAM.

The running time of different algorithms is shown as Table 1. ST and our algorithms are multiscale versions so that longer time is needed to implement. Our proposed algorithm has the highest time complexity. The most time consuming part is iteration of optimization. Reducing computational

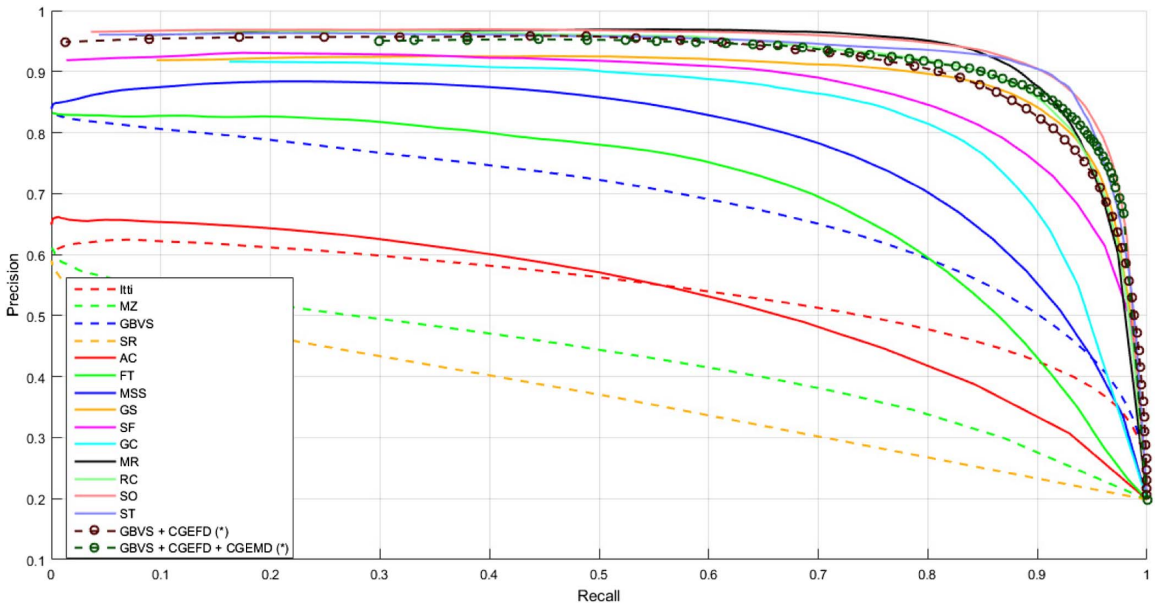

Fig. 9. Quantitative comparison of different salient object detection approaches.

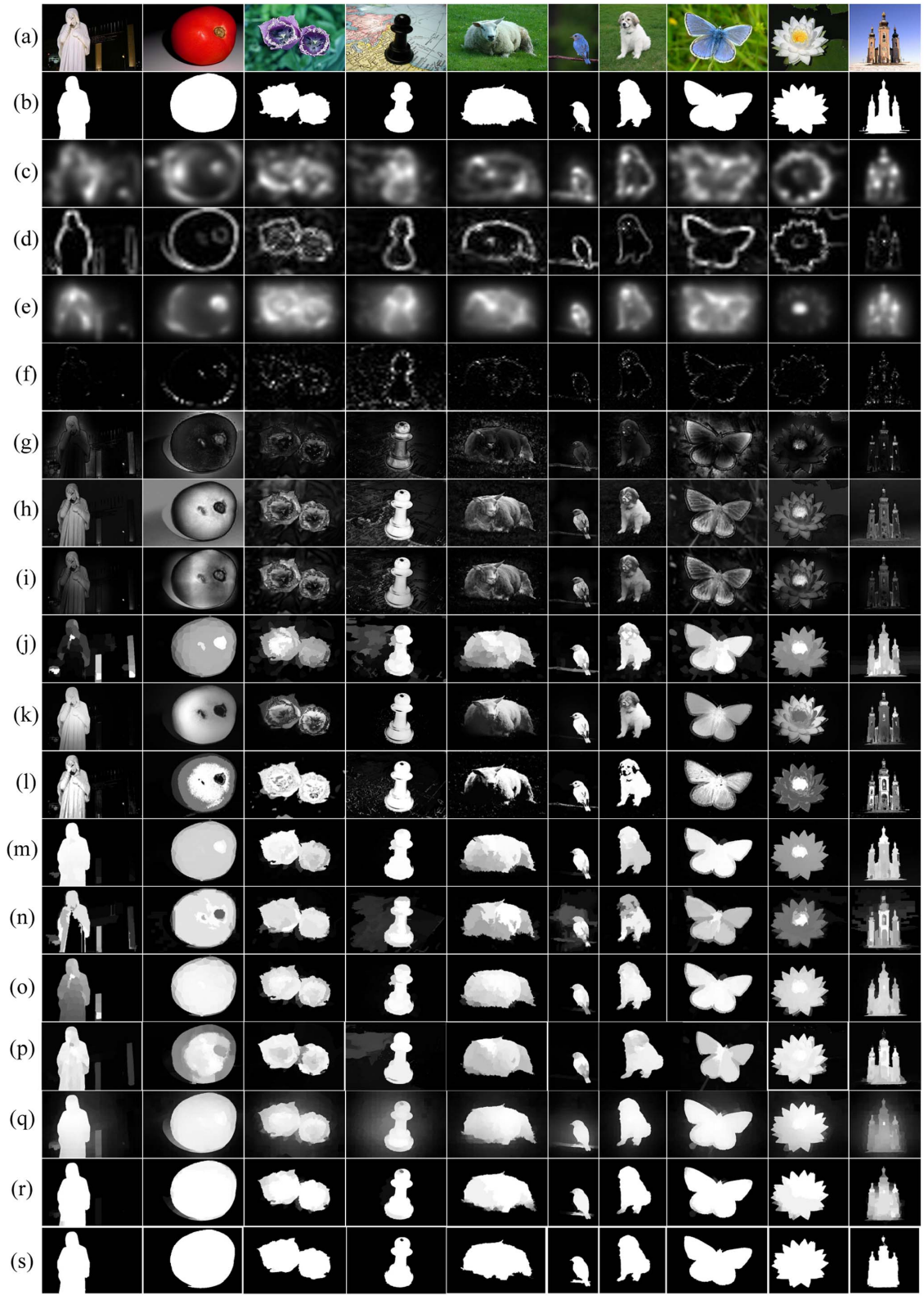

Fig. 10. Salient object detection results of different approaches. (a) Original test images (available on http://research.microsoft.com/en-us/um/people/jiansun/SalientObject/salient_object.htm); (b) ground-truth of the test images (available on http://ivrlwww.epfl.ch/supplementary_material/RK_CVPR09/); (c) saliency detection result from Itti *et al.* [8]; (d) saliency detection result of MZ[9]; (e) saliency detection result of GBVS [10]; (f) saliency detection result of SR [48]; (g) saliency detection result of AC [48]; (h) saliency detection result of FT [12]; (i) saliency detection result of MSS [49]; (j) saliency detection result of GS [15]; (k) saliency detection result of SF [14]; (l) saliency detection result of GC [50]; (m) saliency detection result of MR [16]; (n) saliency detection result of RC [51]; (o) saliency detection result of SO [17]; (p) saliency detection result of ST [45]; (q) saliency detection result of the proposed GBVS + CGEFD; (r) saliency detection result of the proposed GBVS + CGEFD + CGEMD; (s) segmented result of (r).

Table 1. Average Running Time to Process an Image (s)

Algorithm	Running Time	Algorithm	Running Time
Itti [8]	0.345	SF [14]	0.301
GBVS [10]	0.815	GS [15]	0.025
AC [48]	0.375	MR [16]	0.198
FT [12]	0.095	SO [17]	0.084
RC [51]	0.141	MSS [49]	0.081
ST [45]	2.012	GC [50]	0.041
Ours	7.986		

complexity and accepting unsupervised feature learning [53] are two of our future research issues.

5. CONCLUSIONS

Inspired by the attention spread of sensory enhancement theory, we propose a Cauchy graph based diffusion model to propagate the saliency. Then we propose a salient object detection framework by utilizing the diffusion model. The initial spatial attention, diffusion model, and multiscale perception grouping are the three key steps to obtain a good salient map for salient object detection. Experimental results show that our diffusion model for attention spread is critical to transfer spatial attention to object attention, which can improve the performance of salient object prediction.

Funding. Aviation Science Foundation of China (20135179007); National Natural Science Foundation of China (NSFC) (41371339).

REFERENCES

1. J. Scholl, "Objects and attention: the state of the art," Cognition **80**, 1–46 (2001).
2. J. Duncan, "Selective attention and the organization of visual information," J. Exp. Psychol. **113**, 501–517 (1984).
3. Z. Chen, "Object-based attention: a tutorial review," Atten. Percept. Psychophys. **74**, 784–802 (2012).
4. A. Borji, D. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: a comparative study," IEEE Trans. Image Process. **22**, 55–69 (2012).
5. T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," Technical Report (MIT, 2012).
6. A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: a benchmark," in *Proceeding of European Conference on Computer Vision* (ECCV, 2012), pp. 414–429.
7. Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille, "The secrets of salient object segmentation," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2014), pp. 280–287.
8. L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," IEEE Trans. Pattern Anal. Mach. Intell. **20**, 1254–1259 (1998).
9. Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proceedings of ACM International Conference on Multimedia* (ACM, 2003), pp. 374–381.
10. J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *Proceedings of Advances in Neural Information Processing Systems* (2006), pp. 545–552.
11. J. Zhang and S. Sclaroff, "Saliency detection: a boolean map approach," in *Proceedings of IEEE International Conference on Computer Vision* (IEEE, 2013), pp. 153–160.
12. R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2009), pp. 1597–1604.
13. M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu, "Global contrast based on salient region detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2011), pp. 409–416.
14. F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: contrast based filtering for salient region detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2012), pp. 733–740.
15. Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proceedings of European Conference on Computer Vision* (ECCV, 2012), pp. 29–42.
16. C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of IEEE Conference on Computer vision and Pattern Recognition* (IEEE, 2013), pp. 3166–3173.
17. W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of IEEE Conference on Computer vision and Pattern Recognition* (IEEE, 2014), pp. 2814–2821.
18. A. Wanning, L. Stanisor, and P. Roelfsema, "Automatic spread of attentional response modulation along Gestalt criteria in primary visual cortex," Nat. Neurosci. **14**, 1243–1244 (2011).
19. J. Poort, F. Raudies, A. Wanning, V. A. Lamme, H. Neumann, and P. R. Roelfsema, "The role of attention in figure-ground segregation in areas V1 and V4 of the visual cortex," Neuron **75**, 143–156 (2012).
20. P. Roelfsema, V. Lamme, and H. Spekreijse, "Object-based attention in the primary visual cortex of the macaque monkey," Nature **395**, 376–381 (1998).
21. J. Tang, G.-J. Qi, M. Wang, and X.-S. Hua, "Video semantic analysis based on structure-senstive anisotropic manifold ranking," Signal Process. **89**, 2313–2323 (2009).
22. R. Liu, J. Cao, Z. Lin, and S. Shan, "Adaptive partial differential equation learning for visual saliency detection," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2014), pp. 3866–3873.
23. B. Jiang, L. Zhang, H. Lu, C. Yang, and M. Yang, "Saliency detection via absorbing Markov chain," in *Proceedings of International Conference on Computer Vision* (ICCV, 2013), pp. 1665–1673.
24. Z. Ren, Y. Hu, L.-T. Chia, and D. Rajan, "Improved saliency detection based on superpixel clustering and saliency propagation," in *Proceedings of ACM Conference on Multimedia* (ACM, 2010), pp. 1099–1102.
25. C. Gong, D. Tao, W. Liu, S. J. Maybank, M. Fang, K. Fu, and J. Yang, "Saliency propagation from simple to difficult," in *Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition* (IEEE, 2015), pp. 2531–2539.
26. D. Luo, C. Ding, F. Nie, and H. Huang, "Cauchy graph embedding," in *Proceedings of International Conference on Machine Learning* (IEEE, 2011), pp. 553–560.
27. I. Jolliffe, *Principal Component Analysis* (Wiley, 2002).
28. T. Cox and M. Cox, *Multidimensional Scaling* (Chapman & Hall 2001).
29. J. Tenenbaum, V. Silva, and J. Langford, "A global geometric framework for nonlinear dimensionality reduction," Science **290**, 2319–2323 (2000).
30. S. Roweis and L. Saul, "Nonlinear dimensionality reduction by locally linear embedding," Science **290**, 2323–2326 (2000).
31. M. Belkin and P. Niyogi, "Laplacian eigenmaps for dimensionality reduction and data representation," Neural Comput. **15**, 1373–1396 (2003).
32. D. Zhou, O. Bousquet, T. Lal, J. Weston, and B. Scholkopf, "Learning with local and global consistency," in *Proceedings of Advances in Neural Information Processing Systems* (2004), pp. 321–328.
33. X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised learning using Gaussian fields and harmonic functions," in *Proceedings of International Conference on Machine Learning* (2003), pp. 912–919.
34. F. Nie, D. Xu, I. Tsang, and C. Zhang, "Flexible manifold embedding: a framework for semi-supervised and unsupervised dimension reduction," IEEE Trans. Image Process. **19**, 1921–1932 (2010).

35. Y. Li, Y. Tan, J. Deng, Q. Wen, and J. Tian, "Cauchy graph embedding optimization for built-up areas detection from high-resolution remote sensing images," IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **8**, 2078–2096 (2015).
36. D. Zhou and B. Schölkopf, "A regularization framework for learning from graph data," in *Proceedings of ICML Workshop on Statistical Relational Learning and Its Connections to Other Fields* (IMLS, 2004), pp. 132–137.
37. D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Scholkopf, "Ranking on data manifolds," in *Proceedings of Advances in Neural Information Processing Systems* (2004), pp. 169–176.
38. N. Qian, "On the momentum term in gradient descent learning algorithms," Neural Networks **12**, 145–151 (1999).
39. R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Ssstrunk, "SLIC superpixels compared to state-of-the-art superpixels methods," IEEE Trans. Pattern Anal. Mach. Intell. **34**, 2274–2282 (2012).
40. Y. Karklin and M. Lewicki, "Emergence of complex cell properties by learning to generalize in natural scenes," Nature **457**, 83–86 (2009).
41. E. Erdem and A. Erdem, "Visual saliency estimation by non-linearity integrating features using region covariance," J. Vis. **13**(4), 11 (2013).
42. Y. Li, Y. Tan, J.-G. Yu, S. Qi, and J. Tian, "Kernel regression in mixed feature spaces for spatio-temporal saliency detection," Comput. Vis. Image Underst. **135**, 126–140 (2015).
43. J.-G. Yu, J. Zhao, J. Tian, and Y. Tan, "Maximal entropy random walk for region-based visual saliency," IEEE Trans. Cybern. **44**, 1661–1672 (2014).
44. Y. Li, Y. Tan, Y. Li, S. Qi, and J. Tian, "Built-up area detection from satellite images using multikernel learning, multifield integrating, and multihypothesis voting," IEEE Geosci. Remote Sens. Lett. **12**, 434–438 (2015).
45. Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: a novel saliency detection framework," IEEE Trans. Image Process. **23**, 1937–1952 (2014).
46. T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2007), pp. 1–8.
47. X. Hou and L. Zhang, "Saliency detection: a spectral residual approach," in *Proceedings of International Conference on Computer Vision* (IEEE, 2007), pp. 1–8.
48. R. Achanta, F. Estrada, P. Wils, and S. Susstrunk, "Salient region detection and segmentation," Lect. Notes Comput. Sci. **5008**, 66–75 (2008).
49. R. Achanta and S. Susstrunk, "Saliency detection using maximum symmetric surround," in *Proceedings of International Conference on Image Processing* (IEEE, 2010), pp. 2653–2656.
50. M. Cheng, J. Warrell, W. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *Proceedings of International Conference on Computer Vision* (IEEE, 2013), pp. 1529–1536.
51. M. Cheng, N. Mitra, X. Huang, P. Torr, and S. Hu, "Global contrast based salient region detection," IEEE Trans. Pattern Anal. Mach. Intell. **37**, 569–582 (2015).
52. https://github.com/elvintan/SalientObjectDetection.
53. Y. Li, C. Tao, Y. Tan, K. Shang, and J. Tian, "Unsupervised multilayer feature learning for satellite image scene classification," IEEE Geosci. Remote Sens. Lett. **13**, 157–161 (2016).