Saliency Propagation from Simple to Difficult

Chen Gong[1,2], Dacheng Tao[2], Wei Liu[3], S.J. Maybank[4], Meng Fang[2], Keren Fu[1], and Jie Yang[1]

[1]Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University
[2]The Centre for Quantum Computation & Intelligent Systems, University of Technology, Sydney
[3]IBM T. J. Watson Research Center
[4]Birkbeck College, London
Please contact: jieyang@sjtu.edu.cn; dacheng.tao@gmail.com

June 9, 2016

Abstract

Saliency propagation has been widely adopted for identifying the most attractive object in an image. The propagation sequence generated by existing saliency detection methods is governed by the spatial relationships of image regions, *i.e.*, the saliency value is transmitted between two adjacent regions. However, for the inhomogeneous difficult adjacent regions, such a sequence may incur wrong propagations. In this paper, we attempt to manipulate the propagation sequence for optimizing the propagation quality. Intuitively, we postpone the propagations to difficult regions and meanwhile advance the propagations to less ambiguous simple regions. Inspired by the theoretical results in educational psychology, a novel propagation algorithm employing the teaching-to-learn and learning-to-teach strategies is proposed to explicitly improve the propagation quality. In the teaching-to-learn step, a teacher is designed to arrange the regions from simple to difficult and then assign the simplest regions to the learner. In the learning-to-teach step, the learner delivers its learning confidence to the teacher to assist the teacher to choose the subsequent simple regions. Due to the interactions between the teacher and learner, the uncertainty of original difficult regions is gradually reduced, yielding manifest salient objects with optimized background suppression. Extensive experimental results on benchmark saliency datasets demonstrate the superiority of the proposed algorithm over twelve representative saliency detectors.

1 Introduction

Saliency detection has attracted intensive attention and achieved considerable progress during the past two decades. Up to now, a great number of detectors based on computational intelligence have been proposed. They can be roughly divided into two categories: *bottom-up* methods that are data and stimulus driven, and *top-down* methods that are task and knowledge driven.

Top-down methods are usually related to the subsequent applications. For example, Maybank [21] proposed a probabilistic definition of salient image regions for image matching. Yang *et al.* [28] combined dictionary learning and Conditional Random Fields (CRFs) to generate discriminative representation of target-specific objects.

Different from top-down methods, bottom-up methods use low-level cues, such as contrast and spectral information, to recognize the most salient regions without realizing content or specific prior knowledge about the targets. The representatives include [11, 10, 16, 4, 22, 19, 14, 17, 8].

Recently, propagation methods have gained much popularity in bottom-up saliency detection and achieved state-of-the-art performance. To conduct saliency propagations, an input image is represented by a graph over the segmented superpixels, in which the adjacent superpixels in the image are connected by weighted edges. The saliency values are then iteratively diffused along these edges from the labeled superpixels to their unlabeled neighbors. However, such propagations may incur errors if the unlabeled adjacent superpixels are inhomogeneous or very dissimilar to the labeled ones. For example, [9] and [12] formulate the saliency propagation process as random walks on the graph. [27] conduct the propagations by employing manifold based diffusion [29]. All these methods generate similar propagation sequences which are heavily influenced by the superpixels' spatial relationships. However, once encountering the

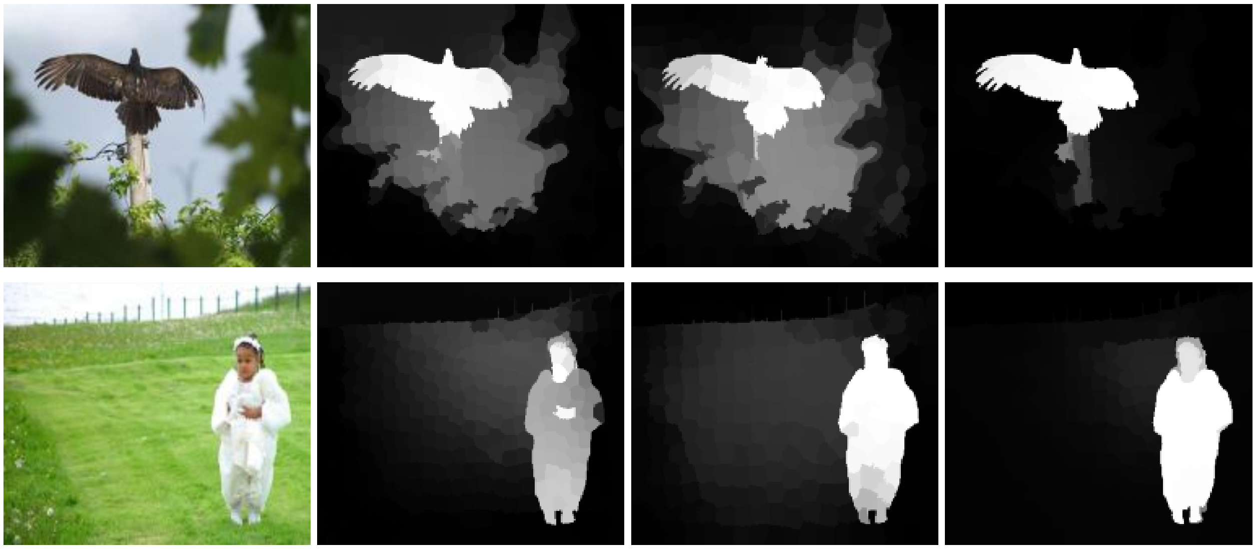

Figure 1: The results achieved by typical propagation methods and our method on two example images. From left to right: input images, results of [27], [12], and our method.

inhomogeneous or incoherent adjacent superpixels, the propagation sequences are misleading and likely to lead to inaccurate detection results (see Fig. 1).

Based on the above observations, we argue that not all neighbors are suitable to participate in the propagation process, especially when they are inhomogeneous or visually different from the labeled superpixels. Therefore, we assume different superpixels have different difficulties, and measure the saliency values of the simple superpixels prior to the difficult ones. This modification to the traditional scheme of generating propagation sequences is very critical, because in this modification the previously attained knowledge can ease the learning burden associated with complex superpixels afterwards, so that the difficult regions can be precisely discovered. Such a "starting simple" strategy conforms to the widely acknowledged theoretical results in pedagogic and cognitive areas [5, 23, 13], which emphasize the importance of teachers for human's acquisitions of knowledge from the childish stage to the mature stage.

By taking advantage of these psychological opinions, we propose a novel approach for saliency propagation by leveraging a teaching-to-learn and learning-to-teach paradigm (displayed in Fig. 3). This paradigm plays two key roles: a teacher behaving as a superpixel selection procedure, and a learner working as a saliency propagation procedure. In the teaching-to-learn step of the t-th propagation, the teacher assigns the simplest superpixels (*i.e.*, curriculum) to the learner in order to avoid the erroneous propagations to the difficult regions. The *informativity*, *individuality*, *inhomogeneity*, and *connectivity* of the candidate superpixels are comprehensively evaluated by the teacher to decide the proper curriculum. In the learning-to-teach step, the learner reports its t-th performance to the teacher in order to assist the teacher in wisely deciding the $(t + 1)$-th curriculum. If the performance is satisfactory, the teacher will choose more superpixels into the curriculum for the following learning process. Owing to the interactions between the teacher and learner, the superpixels are logically propagated from simple to difficult with the updated curriculum, resulting in more confident and accurate saliency maps than those of typical methods (see Fig. 1).

2 Saliency Detection Algorithm

This section details our saliency detection scheme (see Fig. 2). When an input image is given, it is pre-processed by computing the convex hull, segmenting into superpixels, and constructing the graph over these superpixels. After that, the saliency values are propagated from the background seeds to form a coarse map (Stage 1). Finally, this map is refined by propagating the saliency information from the most confident foreground regions to the remaining superpixels (Stage 2). In the above stages, all the propagations are implemented under the proposed teaching-to-learn and learning-to-teach paradigm (see the magenta arrows in Fig. 2), which will be concretely introduced in Section 3.

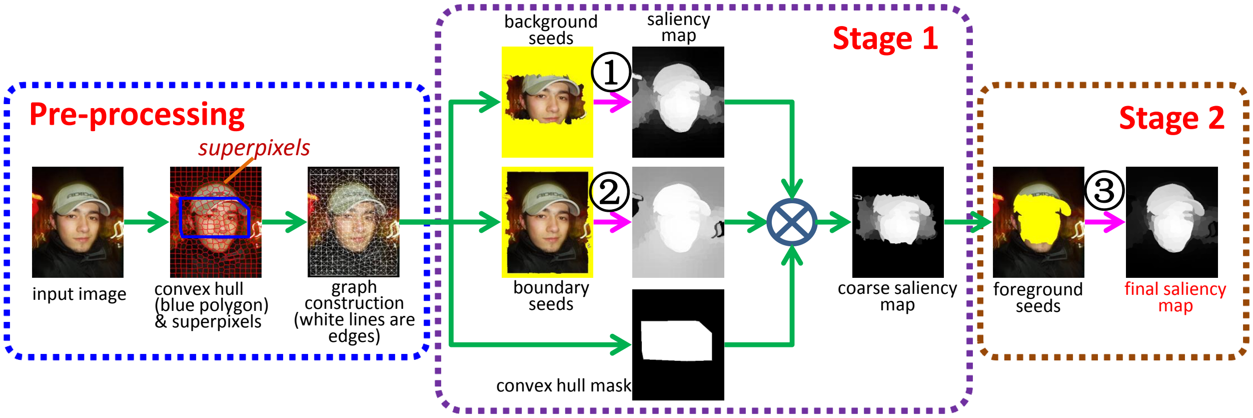

Figure 2: The diagram of our detection algorithm. The magenta arrows annotated with numbers denote the implementations of teaching-to-learn and learning-to-teach propagation shown in Fig. 3.

2.1 Image Pre-processing

Given an input image, a convex hull $\mathcal{H}$ is constructed to estimate the target's location [26]. This is done by detecting some key points in the image via Harris corner detector. Because most key points locate within the target region, we link the outer key points to a convex hull to roughly enclose the target (see Fig. 2).

We proceed by using the SLIC [1] algorithm to over-segment the input image into N small superpixels (see Fig. 2), then an undirected graph $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$ is built where $\mathcal{V}$ is the node set consisted of these superpixels and $\mathcal{E}$ is the edge set encoding the similarity between them. In our work, we link two nodes[1] $\mathbf{s}_i$ and $\mathbf{s}_j$ by an edge if they are spatially adjacent in the image or both of them correspond to the boundary superpixels. Then their similarity is computed by the Gaussian kernel function $\omega_{ij} = \exp\left(-\|\mathbf{s}_i - \mathbf{s}_j\|^2/(2\theta^2)\right)$, where θ is the kernel width and $\mathbf{s}_i$ is the feature vector of the i-th superpixel represented in the LAB-XY space (*i.e.* $\mathbf{s}_i = (\mathbf{s}_i^{color}; \mathbf{s}_i^{position})$). Therefore, the $\mathcal{G}$'s associated adjacency matrix $\mathbf{W} \in \mathbb{R}^{N\times N}$ is defined by $\mathbf{W}_{ij} = \omega_{ij}$ if $i \neq j$, and $\mathbf{W}_{ij} = 0$ otherwise. The diagonal degree matrix is $\mathbf{D}$ with $\mathbf{D}_{ii} = \sum_j \mathbf{W}_{ij}$.

2.2 Coarse Map Establishment

A coarse saliency map is built from the perspective of background, to assess how these superpixels are distinct from the background. To this end, some regions that are probably background should be determined as seeds for the saliency propagation. Two background priors are adopted to initialize the background propagations. The first one is the *convex hull* prior [26] that assumes the pixels outside the convex hull are very likely to be the background; and the second one is the *boundary prior* [24, 27] which indicates the regions along the image's four boundaries are usually non-salient.

For employing the convex hull prior, the superpixels outside $\mathcal{H}$ are regarded as background seeds (marked with yellow in Fig. 2) for saliency propagation. Suppose the propagation result is expressed by an N-dimensional vector $\mathbf{f}^* = \begin{pmatrix} f_1^* & \cdots & f_N^* \end{pmatrix}^T$, where f_i^* $(i = 1, \cdots, N)$ are obtained saliency values corresponding to the superpixels $\mathbf{s}_i$, then after scaling $\mathbf{f}^*$ to $[0, 1]$ (denoted as $\mathbf{f}^*_{normalized}$), the value of the i-th superpixel in the saliency map $S_{ConvexHull}$ is

$$S_{ConvexHull}(i) = 1 - \mathbf{f}^*_{normalized}(i), i = 1, 2, \cdots, N, \tag{2.1}$$

Similarly, we treat the superpixels of four boundaries as seeds, and implement the propagation again. A saliency map based on the boundary prior can then be generated, which is denoted as $S_{Boundary}$. Furthermore, we establish a binary mask S_{mask} [7] to indicate whether the i-th superpixel is inside ($S_{Mask}(i) = 1$) or outside ($S_{Mask}(i) = 0$) the convex hull $\mathcal{H}$. Finally, the saliency map of Stage 1 is obtained by integrating $S_{ConvexHull}$, $S_{Boundary}$, and S_{Mask} as

$$S_{Stage1} = S_{ConvexHull} \otimes S_{Boundary} \otimes S_{Mask}, \tag{2.2}$$

where "$\otimes$" is the element-wise product between matrices.

[1] In this paper, "superpixel" and "node" refer to the same thing. We use them interchangeably for different explanation purposes.

2.3 Map Refinement

After the Stage 1, the dominant object can be roughly highlighted. However, S_{Stage1} may still contain some background noise that should be suppressed. Therefore, we need to propagate the saliency information from the potential foreground regions to further improve S_{Stage1}.

Intuitively, we may choose the superpixels with large saliency values in S_{Stage1} as foreground seeds. In order to avoid erroneously taking background as seeds, we carefully pick up a small number of superpixels as seeds that are in the set:

$$\{\mathbf{s}_i|\ S_{Stage1}(i) \geq \eta \mathrm{max}_{1\leq j\leq N}(S_{Stage1}(j))\}\,, \tag{2.3}$$

where η is set to 0.7. Finally, by setting the labels of seeds to 1 and conducting the teaching-to-learn and learning-to-teach propagation, we achieve the final saliency map S_{Stage2}. Fig. 2 illustrates that S_{Stage2} successfully highlights the foreground regions while removes the background noise appeared in S_{Stage1}.

3 Teaching-to-learn and Learning-to-teach For Saliency Propagation

Saliency propagation plays an important role in our algorithm. Suppose we have l seed nodes $\mathbf{s}_1, \cdots, \mathbf{s}_l$ on $\mathcal{G}$ with saliency values $f_1 = \cdots = f_l = 1$, the task of saliency propagation is to reliably and accurately transmit these values from the l labeled nodes to the remaining $u = N - l$ unlabeled superpixels.

As mentioned in the introduction, the propagation sequence in existing methods [9, 27, 12] may incur imperfect results on difficult superpixels, so we propose a novel teaching-to-learn and learning-to-teach framework to optimize the learning sequence (see Fig. 3). To be specific, this framework consists of a learner and a teacher. Given the labeled set and unlabeled set at time t denoted as $\mathcal{L}^{(t)}$ and $\mathcal{U}^{(t)}$, the teacher selects a set of simple superpixels from $\mathcal{U}^{(t)}$ as curriculum $\mathcal{T}^{(t)}$. Then, the learner will learn $\mathcal{T}^{(t)}$, and return a feedback to the teacher to help the teacher update the curriculum for the $(t+1)$-th learning. This process iterates until all the superpixels in $\mathcal{U}^{(t)}$ are properly learned.

3.1 Teaching-to-learn

The core of teaching-to-learn is to design a teacher deciding which unlabeled superpixels are to be learned. For the t-th propagation, a candidate set $\mathcal{C}^{(t)}$ is firstly established, in which the elements are nodes directly connected to the labeled set $\mathcal{L}^{(t)}$ on $\mathcal{G}$. Then the teacher chooses the simplest superpixels from $\mathcal{C}^{(t)}$ as the t-th curriculum. To evaluate the propagation difficulty of an unlabeled superpixel $\mathbf{s}_i \in \mathcal{C}^{(t)}$, the difficulty score DS_i is defined by combining informativity INF_i, individuality IND_i, inhomogeneity IHM_i, and connectivity CON_i, namely:

$$DS_i = INF_i + \beta_1 IND_i + \beta_2 IHM_i + \beta_3 CON_i, \tag{3.1}$$

where β_1, β_2 and β_3 are weighting parameters. Next we will detail the definitions and computations of INF_i, IND_i, IHM_i, and CON_i, respectively.

Informativity: The simple superpixel should not contain too much information given the labeled set $\mathcal{L}$[2]. Therefore, the informativity of a superpixel $\mathbf{s}_i \in \mathcal{C}$ is straightforwardly modelled by the conditional entropy $H(\mathbf{s}_i|\mathcal{L})$, namely:

$$INF_i = H(\mathbf{s}_i|\mathcal{L}). \tag{3.2}$$

The propagations on the graph follow the multivariate Gaussian process [31], with the elements f_i $(i = 1, \cdots, N)$ in the random vector $\mathbf{f} = \begin{pmatrix} f_1 & \cdots & f_N \end{pmatrix}^T$ denoting the saliency values of superpixels $\mathbf{s}_i$. The associated covariance matrix $\mathbf{K}$ equals to the adjacency matrix $\mathbf{W}$ except the diagonal elements are set to 1.

For the multivariate Gaussian, the closed-form solution of $H(\mathbf{s}_i|\mathcal{L})$ is [3]:

$$H(\mathbf{s}_i|\mathcal{L}) = \frac{1}{2}\ln(2\pi e \sigma^2_{i|\mathcal{L}}), \tag{3.3}$$

where $\sigma^2_{i|\mathcal{L}}$ denotes the conditional covariance of f_i given $\mathcal{L}$. Considering that the conditional distribution is a multivariate Gaussian, $\sigma^2_{i|\mathcal{L}}$ in (3.3) can be represented by

$$\sigma^2_{i|\mathcal{L}} = \mathbf{K}^2_{ii} - \mathbf{K}_{i,\mathcal{L}}\mathbf{K}^{-1}_{\mathcal{L},\mathcal{L}}\mathbf{K}_{\mathcal{L},i}, \tag{3.4}$$

in which $\mathbf{K}_{i,\mathcal{L}}$ and $\mathbf{K}_{\mathcal{L},\mathcal{L}}$ denote the sub-matrices of $\mathbf{K}$ indexed by the corresponding subscripts. By plugging (3.3) and (3.4) into (3.2), we obtain the informativity of $\mathbf{s}_i$.

[2] For simplicity, the superscript t is omitted for all the notations hereinafter unless otherwise specified.

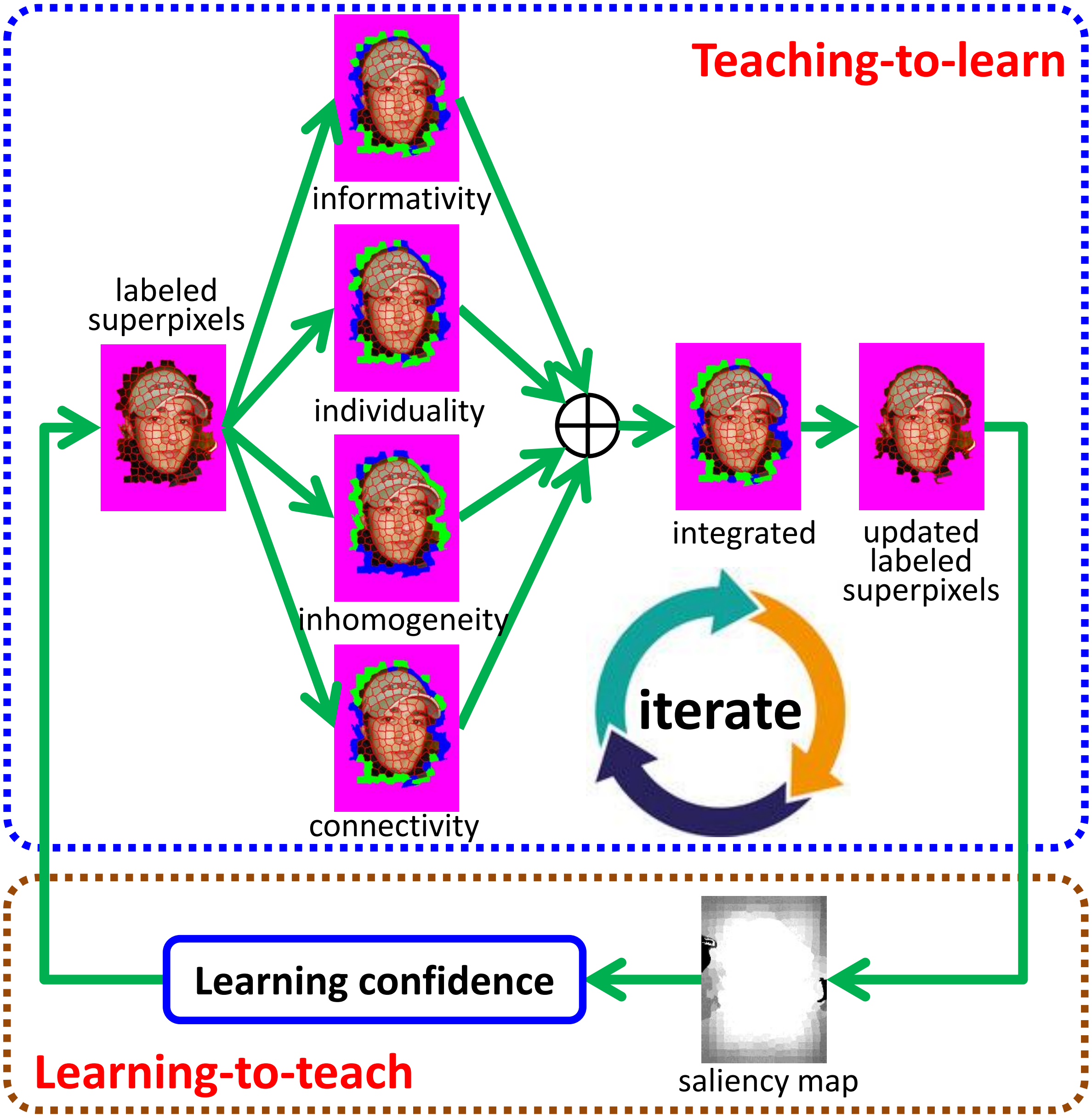

Figure 3: An illustration of our teaching-to-learn and learning-to-teach paradigm. In the teaching-to-learn step, based on a set of labeled superpixels (magenta) in an image, the teacher discriminates the adjacent unlabeled superpixels as difficult (blue superpixels) or simple (green superpixels) by fusing their informativity, individuality, inhomogeneity, and connectivity. Then simple superpixels are learned by the learner, and the labeled set is updated correspondingly. In the learning-to-teach step, the learner provides a learning feedback to the teacher to help decide the next curriculum.

In (3.4), the inverse of an $l \times l$ (l is the size of gradually expanded labeled set $\mathcal{L}$) matrix $\mathbf{K}_{\mathcal{L},\mathcal{L}}$ should be computed in every iteration. As l becomes larger and larger, directly inverting this matrix can be time-consuming. Therefore, an efficient updating technique is developed in the **supplementary material** based on the blockwise inversion equation.

Individuality: Individuality measures how distinct of a superpixel to its surrounding superpixels. We consider a superpixel simple if it is similar to the nearby superpixels in the LAB color space. This is because such superpixel is very likely to share the similar saliency value with its neighbors, thus can be easily identified as either foreground or background. For example, the superpixel $\mathbf{s}_2$ in Fig. 4(a) has lower individuality than $\mathbf{s}_1$ since it is more similar to the neighbors than $\mathbf{s}_1$. The equation below quantifies the local individuality of $\mathbf{s}_i$ and its neighboring superpixels $\mathcal{N}(\mathbf{s}_i)$:

$$IND_i = IND(\mathbf{s}_i, \mathcal{N}(\mathbf{s}_i)) = \frac{1}{|\mathcal{N}(\mathbf{s}_i)|} \sum_{j \in \mathcal{N}(\mathbf{s}_i)} \left\| \mathbf{s}_i^{color} - \mathbf{s}_j^{color} \right\|, \tag{3.5}$$

where $|\mathcal{N}(\mathbf{s}_i)|$ denotes the amount of $\mathbf{s}_i$'s neighbors. Consequently, the superpixels with small individuality are preferred for the current learning.

Inhomogeneity: It is obvious that a superpixel is ambiguous if it is not homogenous or compact. Fig. 4(b) provides an example that the homogenous $\mathbf{s}_4$ gets smaller IHM than the complicated $\mathbf{s}_3$. Suppose there are b pixels $\left\{\mathbf{p}_j^{color}\right\}_{j=1}^{b}$ in a superpixel $\mathbf{s}_i$ characterized by the LAB color feature, then their pairwise correlations are recorded in the $b \times b$ symmetric matrix $\mathbf{\Theta} = \mathbf{P}\mathbf{P}^T$, where $\mathbf{P}$ is a matrix with each row representing a pixel $\mathbf{p}_j^{color}$. Therefore, the inhomogeneity of a superpixel $\mathbf{s}_i$ is defined by the reciprocal of mean value of all the pairwise correlations:

$$IHM_i = \left(\frac{2}{b^2 - b} \sum\nolimits_{i=1}^{b} \sum\nolimits_{j=i+1}^{b} \mathbf{\Theta}_{ij} \right)^{-1}, \tag{3.6}$$

where $\mathbf{\Theta}_{ij}$ is the (i,j)-th element of matrix $\mathbf{\Theta}$. Small IHM_i means that all the pixels in $\mathbf{s}_i$ are much correlated with others, so $\mathbf{s}_i$ is homogenous and can be easily learned.

Connectivity: For the established graph $\mathcal{G}$, a simple intuition is that the nodes strongly connected to the labeled set $\mathcal{L}$ are not difficult to propagate. Such strength of connectivity is inversely proportional to the averaged geodesic distances between $\mathbf{s}_i \in \mathcal{C}$ and all the elements in $\mathcal{L}$, namely:

$$CON_i = \frac{1}{l} \sum\nolimits_{j \in \mathcal{L}} geo(\mathbf{s}_i, \mathbf{s}_j). \tag{3.7}$$

In (3.7), $geo(\mathbf{s}_i, \mathbf{s}_j)$ represents the geodesic distance between $\mathbf{s}_i$ and $\mathbf{s}_j$, which can be approximated by their shortest path, namely:

$$\begin{array}{l} geo(\mathbf{s}_i, \mathbf{s}_j) = \min\limits_{R_1=i, R_2, \cdots, R_n=j} \sum\nolimits_{k=1}^{n-1} \max(E_{R_k, R_{k+1}} - c_0, 0) \\ s.t. \quad R_k, R_{k+1} \in \mathcal{V}, \quad R_k \text{ and } R_{k+1} \text{ are connected in } \mathcal{G} \end{array}. \tag{3.8}$$

Here $\mathcal{V}$ denotes the nodes set of $\mathcal{G}$, $E_{R_k,R_{k+1}}$ computes the Euclidean distance between R_k and R_{k+1}, and c_0 is an adaptive threshold preventing the "small-weight-accumulation" problem [24].

Finally, by substituting (3.2), (3.5), (3.6) and (3.7) into (3.1), the difficulty scores of all $\mathbf{s}_i \in \mathcal{C}$ can be calculated, based on which the teacher is able to determine the simple curriculum for the current iteration. With the teacher's effort, the unlabeled superpixels are gradually learned from simple to difficult, which is different from the propagation sequence in many existing methodologies [27, 12, 9]. Suppose there are $|\mathcal{C}|$ superpixels in candidate set $\mathcal{C}$, then the weighting parameters β_1, β_2 and β_3 in (3.1) are decided by

$$\begin{cases} \beta_1 = \mathrm{var}\left(IND_1, \cdots, IND_{|\mathcal{C}|}\right) / \mathrm{var}\left(INF_1, \cdots, INF_{|\mathcal{C}|}\right) \\ \beta_2 = \mathrm{var}\left(IHM_1, \cdots, IHM_{|\mathcal{C}|}\right) / \mathrm{var}\left(INF_1, \cdots, INF_{|\mathcal{C}|}\right), \\ \beta_3 = \mathrm{var}\left(CON_1, \cdots, CON_{|\mathcal{C}|}\right) / \mathrm{var}\left(INF_1, \cdots, INF_{|\mathcal{C}|}\right) \end{cases} \tag{3.9}$$

where $\mathrm{var}(\cdot)$ is the variance computation operator. In (3.9), the metric with large variance is assigned to large weight, because it properly reflects the difference of candidate superpixels.

3.2 Learning-to-teach

After the difficulty scores of all candidate superpixels are computed, the next step is to pick up a certain number of superpixels as curriculum based on $DS_1, \cdots, DS_{|\mathcal{C}|}$. A straightforward idea is to sort all the elements in $\mathcal{C}$ so that their difficulty scores satisfying $DS_1 \le DS_2 \le \cdots \le DS_{|\mathcal{C}|}$. Then the first q ($q \le |\mathcal{C}|$) superpixels are used to establish the curriculum set $\mathcal{T} = \{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_q\}$ according to the pre-defined q. However, we hold that how many superpixels are to

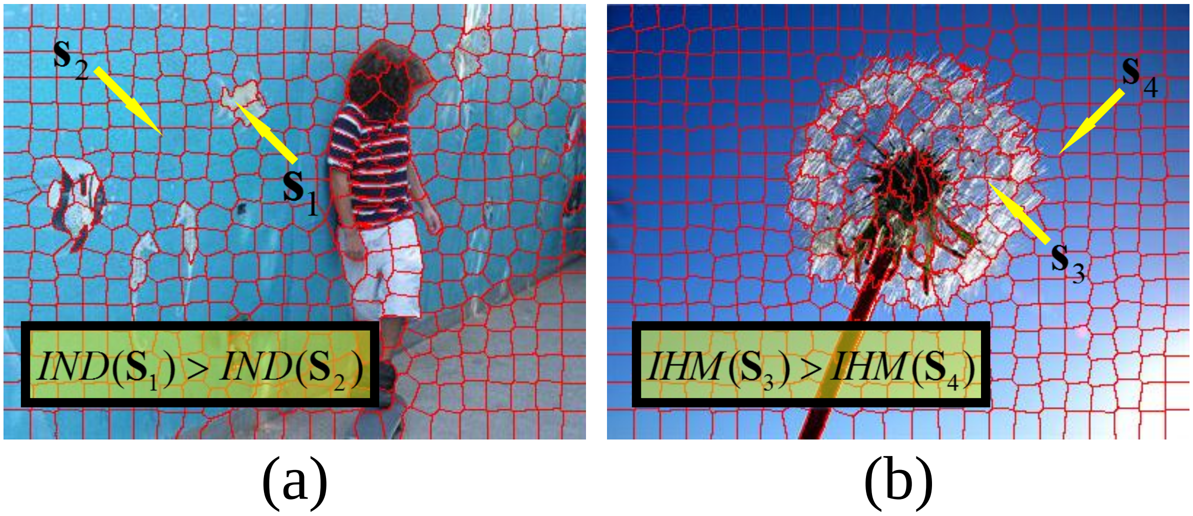
$\mathbf{s}_2$
$\mathbf{s}_1$
$IND(\mathbf{S}_1) > IND(\mathbf{S}_2)$
$\mathbf{s}_4$
$\mathbf{s}_3$
$IHM(\mathbf{S}_3) > IHM(\mathbf{S}_4)$
(a)
(b)

Figure 4: The illustrations of individuality (a) and inhomogeneity (b). The region $\mathbf{s}_1$ in (a) obtains larger individuality than $\mathbf{s}_2$, and $\mathbf{s}_3$ in (b) is more inhomogeneous than $\mathbf{s}_4$.

be learned at t should depend on the $(t-1)$-th learning performance. If the $(t-1)$-th learning is confident, the teacher may assign "heavier" curriculum to the learner. In other words, the teacher should also consider the learner's feedback to arrange the proper curriculum, which is a "learning-to-teach" mechanism. Next we will use this mechanism to adaptively decide $q^{(t)}$ for the t-th curriculum.

As mentioned above, $q^{(t)}$ should be adjusted by considering the effect of previous learning. However, since the correctness of the $(t-1)$-th output saliency is unknown, we define a confidence score to blindly evaluate the previous learning performance. Intuitively, the $(t-1)$-th learning is confident if the saliency values $f_1^{(t-1)}, \cdots, f_{q^{(t-1)}}^{(t-1)}$ are close to 0 (very dissimilar to seeds) or 1 (very similar to seeds) after scaling. However, if $f_1^{(t-1)}, \cdots, f_{q^{(t-1)}}^{(t-1)}$ are close to the ambiguous value 0.5, the teacher will rate the last learning as unsatisfactory, and produce a small $q^{(t)}$ to relieve the "burden" for the current learning. Therefore, the confidence score that belongs to $[0, 1]$ is defined by

$$ConfidenceScore = 1 - \frac{2}{q^{(t-1)}} \sum\nolimits_{i=1}^{q^{(t-1)}} \min(f_i^{(t-1)}, 1 - f_i^{(t-1)}), \tag{3.10}$$

and $q^{(t)}$ is finally computed by

$$q^{(t)} = \left\lceil \left|\mathcal{C}^t\right| \times ConfidenceScore \right\rceil . \tag{3.11}$$

3.3 Saliency Propagation

After the curriculum $\mathcal{T}^{(t)} = \left\{\mathbf{s}_1, \mathbf{s}_2, \cdots, \mathbf{s}_{q^{(t)}}\right\}$ is specified, the learner will spread the saliency values from $\mathcal{L}^{(t)}$ to $\mathcal{T}^{(t)}$ via propagation. Particularly, the expression is:

$$\mathbf{f}^{(t+1)} = \mathbf{M}^{(t)} \mathbf{D}^{-1} \mathbf{W} \mathbf{f}^{(t)}, \tag{3.12}$$

where $\mathbf{M}^{(t)}$ is a diagonal matrix with $\mathbf{M}_{ii}^{(t)} = 1$ if $\mathbf{s}_i \in \mathcal{L}^{(t)} \cup \mathcal{T}^{(t)}$, and $\mathbf{M}_{ii}^{(t)} = 0$ otherwise. When the t-th iteration is completed, the labeled and unlabeled sets are updated as $\mathcal{L}^{(t+1)} = L^{(t)} \cup \mathcal{T}^{(t)}$ and $\mathcal{U}^{(t+1)} = \mathcal{U}^{(t)} \backslash \mathcal{T}^{(t)}$, respectively. (3.12) initializes from the binary vector $\mathbf{f}^{(0)} = \left(f_1^{(0)}, \cdots, f_N^{(0)}\right)^{\mathcal{T}}$ ($f_i^{(0)} = 1$ if the i-th superpixel corresponds to seed, and 0 otherwise), terminates when $\mathcal{U}$ becomes an empty set, and the obtained saliency value vector is denoted by $\bar{\mathbf{f}}$. Finally, we smooth $\bar{\mathbf{f}}$ by driving the entire propagation on $\mathcal{G}$ to the stationary state:

$$\mathbf{f}^* = \left(\mathbf{I} - \alpha \mathbf{D}^{-1} \mathbf{W}\right)^{-1} \bar{\mathbf{f}}, \tag{3.13}$$

where α is a parameter set to 0.99 [27], and $\mathbf{f}^*$ encodes the saliency information of N superpixels as defined in Section 2.2.

One example of the complete propagation process is visualized in Fig. 5, in which the superpixels along the image's four boundaries serve as seeds to propagate the saliency information to the remaining superpixels (see Fig.

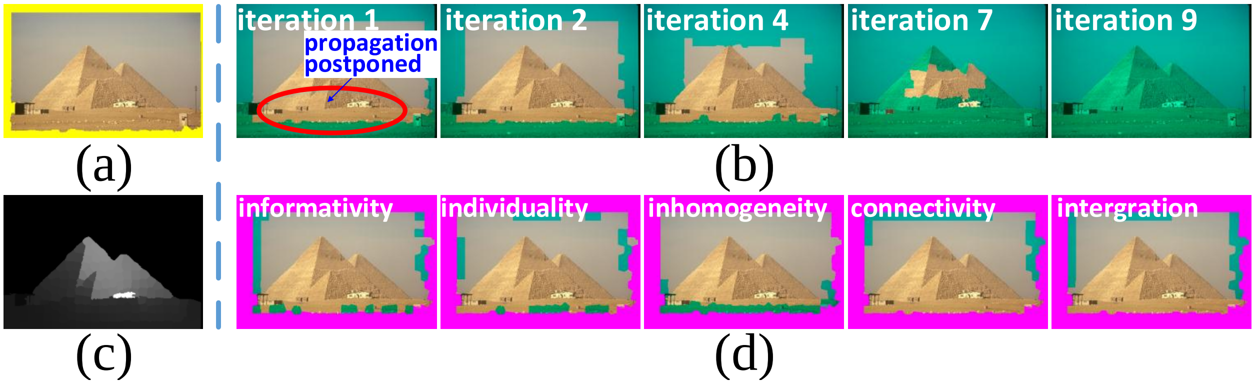

Figure 5: Visualization of the designed propagation process. (a) shows the input image with boundary seeds (yellow). (b) displays the propagations in several key iterations, and the expansions of labeled set $\mathcal{L}$ are highlighted with light green masks. (c) is the final saliency map. The curriculum superpixels of the 2nd iteration decided by informativity, individuality, inhomogeneity, connectivity, and the final integrated result are visualized in (d), in which the magenta patches represent the learned superpixels in the 1st propagation, and the regions for the 2nd diffusion are annotated with light green.

5(a)). In (b), we observe that the sky regions are relatively easy and are firstly learned during the 1st∼4th iterations. In contrast, the land areas are very different from the seeds, so they are difficult and their diffusion should be deferred. Though the labeled set touches the land in a very early time (see the red circle in the 1st iteration), the land superpixels are not diffused until the 4th iteration. This is because the background regions are mostly learned until the 4th iteration, which provide sufficient preliminary knowledge to identify the difficult land regions as foreground or background. As a result, the learner is more confident to assign the correct saliency values to the land after the 4th iteration, and the target (pyramid) is learned in the end during the 7th∼9th iterations. More concretely, the effect of our curriculum selection approach is demonstrated in Fig. 5(d). It can be observed that though the curriculum superpixels are differently chosen by their informativity, individuality, inhomogeneity, and connectivity, they are easy to learn based on the previous accumulated knowledge. Particularly, we notice that the final integrated result only preserves the sky regions for the leaner, while discards the land areas though they are recommended by informativity, individuality, and inhomogeneity. This further reduces the erroneous propagation possibility since the land looks differently from the sky and actually more similar to the unlearned pyramid. Therefore, the fusion scheme (3.1) and the proper $q^{(t)}$ decided by the learning-to-teach step are reasonable and they are critical to the successful propagations (see Fig. 5(c)).

4 Physical Interpretation and Justification

A key factor to the effectiveness of our method is the well-ordered learning sequence from simple to difficult, which is also considered by curriculum learning [2] and self-paced learning [15]. This paper introduces this strategy to graph-based saliency propagation. More interestingly, we provide a physical interpretation of this strategy, by relating the curriculum guided propagation to the practical fluid diffusion.

In physics, *Fick's Law of Diffusion* [6] is well-known for understanding the mass transfer of solids, liquids, and gases through diffusive means. It postulates that the flux diffuses from regions of high concentration to regions of low concentration, with a magnitude that is proportional to the concentration gradient (see Fig. 6(a)). Along one diffusive direction, the law is formulated as

$$J = -\gamma \frac{\partial h}{\partial \delta}, \tag{4.1}$$

where γ is the diffusion coefficient, δ is the diffusion distance, h is the concentration that evaluates the density of molecules of fluid, and J is the diffusion flux that measures the quantity of molecules flowing through the unit area per unit time.

We regard the seed superpixels as sources to emit the fluid, and the remaining unlabeled superpixels are to be diffused, among which the simple and difficult superpixels are compared to lowlands and highlands, respectively (see Figs. 6(b)(c)). There are two obvious facts here: 1) the lowlands will be propagated prior to the highlands, and 2) fluid

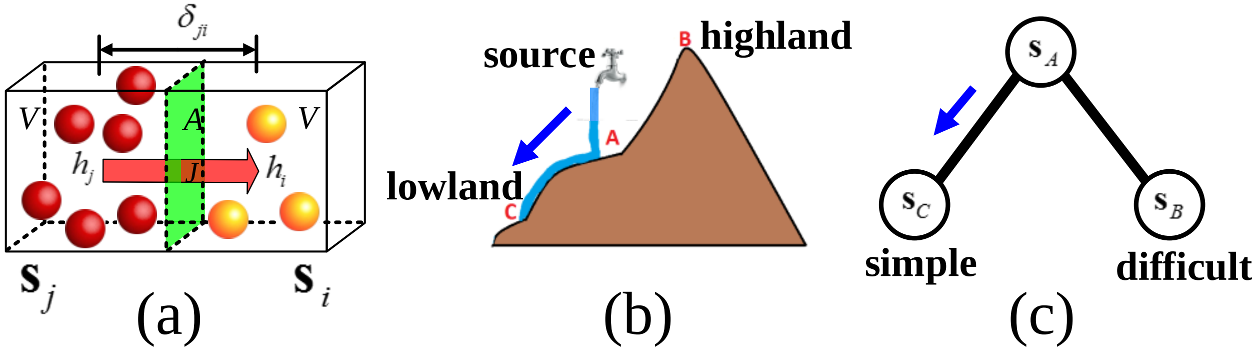

Figure 6: The physical interpretation of our saliency propagation algorithm. (a) analogies the propagation between two regions with equal difficulty to the fluid diffusion between two cubes with same altitude. The left cube with more balls is compared to the region with larger saliency value. The right cube with fewer balls is compared to the region with less saliency cues. The red arrow indicates the diffusion direction. (b) and (c) draw the parallel between fluid diffusion with different altitudes and saliency propagation guided by curriculums. The lowland "C", highland "B", and source "A" in (b) correspond to the simple node $\mathbf{s}_C$, difficult node $\mathbf{s}_B$, and labeled node $\mathbf{s}_A$ in (c), respectively. Like the fluid can only flow from "A" to the lowland "C" in (b), $\mathbf{s}_A$ in (c) also tends to transfer the saliency value to the simple node $\mathbf{s}_C$.

cannot be transmitted from lowlands to highlands. Therefore, by treating γ as the propagation coefficient, h as the saliency value (equivalent to f in above sections), and δ as the propagation distance defined by $\delta_{ji} = 1/\sqrt{\omega_{ji}}$, (4.1) explains the process of saliency propagation from $\mathbf{s}_j$ to $\mathbf{s}_i$ as

$$J_{ji} = -m_i\gamma\frac{f_i^{(t)} - f_j^{(t)}}{\delta_{ji}} = -m_i\gamma\sqrt{\omega_{ji}}(f_i^{(t)} - f_j^{(t)}). \tag{4.2}$$

The parameter m_i in (4.2), which plays the same role as $\mathbf{M}_{ii}$ in (3.12), denotes the "altitude" of $\mathbf{s}_i$. It equals to 1 if $\mathbf{s}_i$ corresponds to a lowland, and 0 if $\mathbf{s}_i$ represents a highland. Note that if $\mathbf{s}_i$ is higher than $\mathbf{s}_j$, the flux $J_{ji} = 0$ because the fluid cannot transfer from lowland to highland. Given (4.2), we have the following theorem:

Theorem 1: Suppose all the superpixels $\mathbf{s}_1, \cdots, \mathbf{s}_N$ in an image are modelled as cubes with volume V, and the area of their interface is A. By using m_i to indicate the altitude of $\mathbf{s}_i$ and setting the propagation coefficient $\gamma = 1$, the proposed saliency propagation can be derived from the fluid transmission modelled by Fick's Law of Diffusion.

We put the detailed proof in the **supplementary material** due to the limited page length. Theorem 1 reveals that our propagation method can be perfectly explained by the well-known physical theory.

5 Experimental Results

In this section, we qualitatively and quantitatively compare the proposed **T**eaching-to-**L**earn and **L**earning-to-**T**each approach (abbreviated as "TLLT") with twelve popular methods on two popular saliency datasets. The twelve baselines include classical methods (LD [18], GS [24]), state-of-the-art methods (SS [8], PD [19], CT [14], RBD [30], HS [25], SF [22]), and representative propagation based methods (MR [27], GP [7], AM [12], GRD [26]). The parameters in our method are set to $N = 400$ and $\theta = 0.25$ throughout the experiments.

5.1 Metrics

Margolin *et al.* [20] point out that the traditional Precision-Recall curve (PR curve) and F_β-measure suffer the interpolation flaw, dependency flaw and equal-importance flaw. Instead, they propose the weighted precision Precision^w, weighted recall Recall^w and weighted F_β-measure F_β^w to achieve more reasonable evaluations. In this paper, we adopt this recently proposed metrics [20] to evaluate the algorithms' performance. The parameter β^2 in $F_\beta^w = (1+\beta^2)\frac{\text{Precision}^w + \text{Recall}^w}{\beta^2\text{Precision}^w + \text{Recall}^w}$ is set to 0.3 as usual to emphasize the precision [27, 25]. Fig. 7 shows some examples that our visually better detection results are underestimated by the existing PR curve, but receive reasonable assessments from the metrics of [20].

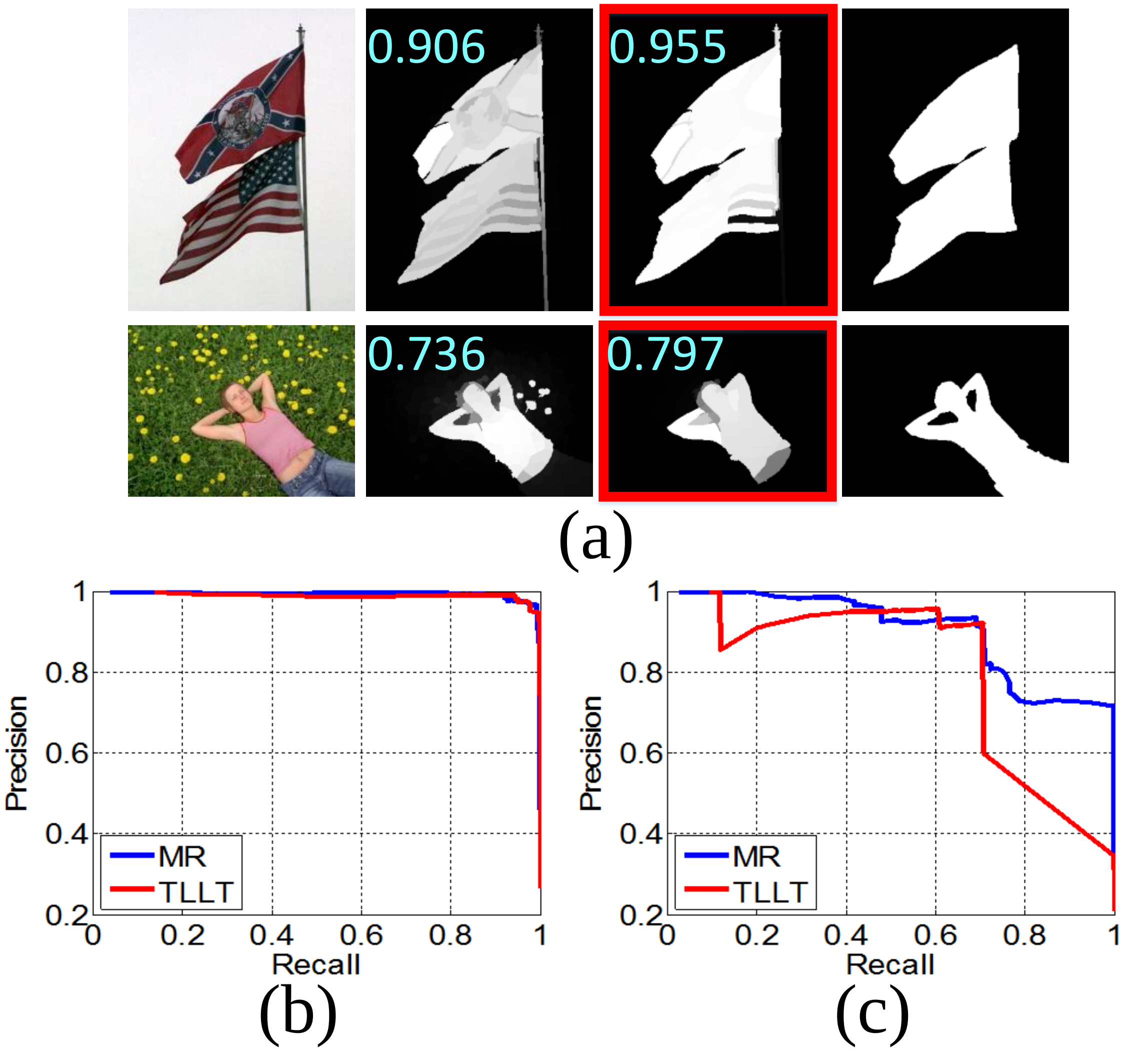

Figure 7: The comparison of traditional PR curve vs. the metric in [20]. (a) shows two saliency maps generated by MR [27] and our method. The columns are (from Left to Right): input images, MR results, our results, and groundtruth. (b), (c) present the PR curves over the images in the first and second rows of (a), respectively. In the top image of (a), our more confident result surprisingly receives the similar evaluation with MR reflected by (b). In the bottom image, the MR result fails to supress the flowers in the background, but turns out to be significantly better than our method revealed by (c). In contrast, the weighted F_β-measure F_β^w (light blue numbers in (a)) provides more reasonable judgements and gives our saliency maps higher evaluations (marked by the red boxes).

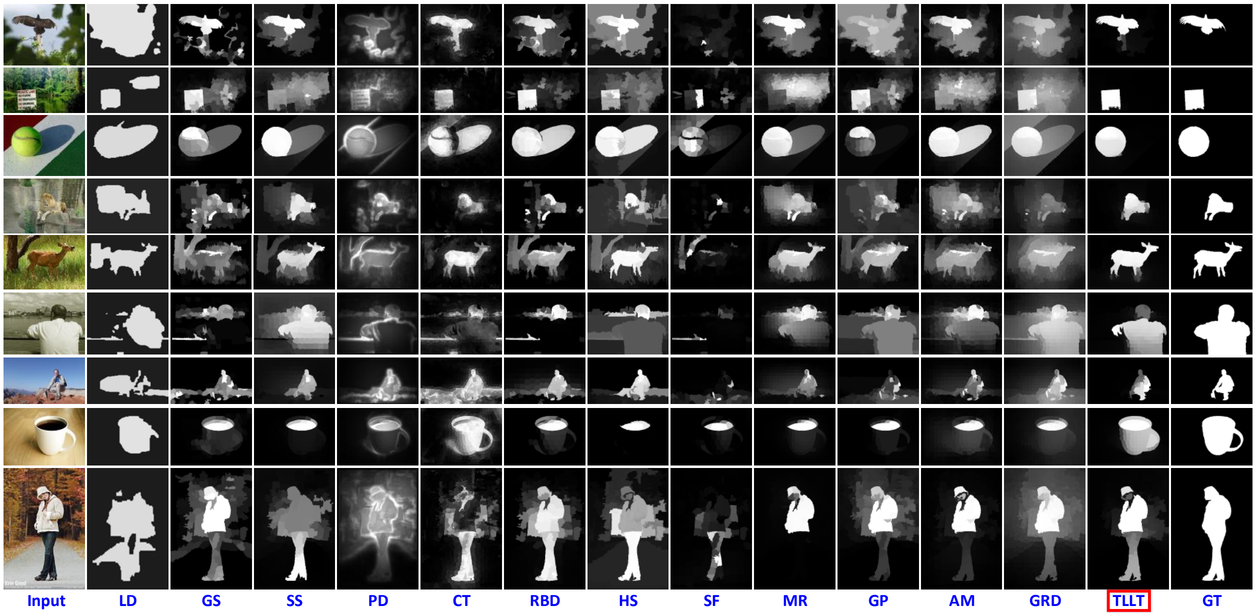

Figure 8: Visual comparisons of saliency maps generated by all the methods on some challenging images. The ground truth (GT) is presented in the last column.

Table 1: Average CPU seconds of all the approaches on ECSSD dataset

Method	LD	GS	SS	PD	CT	RBD	HS	SF	MR	GP	AM	GRD	TLLT
Duration (s)	7.24	0.18	3.58	2.87	3.53	0.20	0.43	0.19	0.87	3.22	0.15	0.93	2.05
Code	matlab	matlab	matlab	matlab	matlab	matlab	C++	matlab	matlab	matlab	matlab	matlab	matlab

5.2 Experiments on Public Datasets

The MSRA 1000 dataset [18], which contains 1000 images with binary pixel-level groundtruth, is firstly adopted for our experiments. The average precision^w, recall^w, and F_β^w of all the methods are illustrated in Fig. 9(a). We can observe that the F_β^w of our TLLT is larger than 0.8, which is the highest record among all the comparators. Another notable fact is that TLLT outperforms other baselines with a large margin in Precision^w. This is because the designed teaching-to-learn and learning-to-teach paradigm propagates the saliency value carefully and accurately. As a result, our approach has less possibility to generate the blurred saliency map with confused foreground. In this way, the Precision^w is significantly improved. More importantly, we note that the Recall^w of our method also touches a relatively high value, although the Precision^w has already obtained an impressive record. This further demonstrates the strength of our innovation.

Although the images from MSRA 1000 dataset have a large variety in their content, the foreground is actually prominent among the simple and structured background. Therefore, a more complicated dataset ECSSD [25], which represents more general situations that natural images fall into, is adopted to further test all the algorithms. Fig. 9(b) shows the result. Generally, all methods perform more poorly on ECSSD than on the MSRA 1000. However, our algorithm still achieves the highest F_β^w and Precision^w when compared with other baselines. RBD obtains slightly lower F_β^w than our method with 0.5215 compared to 0.5283, but the weighted precision is not as good as our approach. Besides, some methods that show very encouraging performance under the traditional PR curve metric, such as HS, SF and GRD, only obtain very moderate results under the new metrics. Since they tend to detect the most salient regions at the expense of low precision, the imbalance between Precision^w and Recall^w will happen, which pulls down the overall F_β^w to a low value. Comparatively, TLLT produces relatively balanced Precision^w and Recall^w on both datasets, therefore higher F_β^w is obtained.

The average CPU seconds of evaluated methods for processing one image in ECSSD are summarized in Tab. 1, on an Intel i5 3.20GHz CPU with 8GB RAM. our method takes 2.05 seconds per detection, which is slower than GS, RBD, HS, SF, MR, AM, GRD, but faster than LD, SS, PD, CT, and GP. Because our method needs to decide the

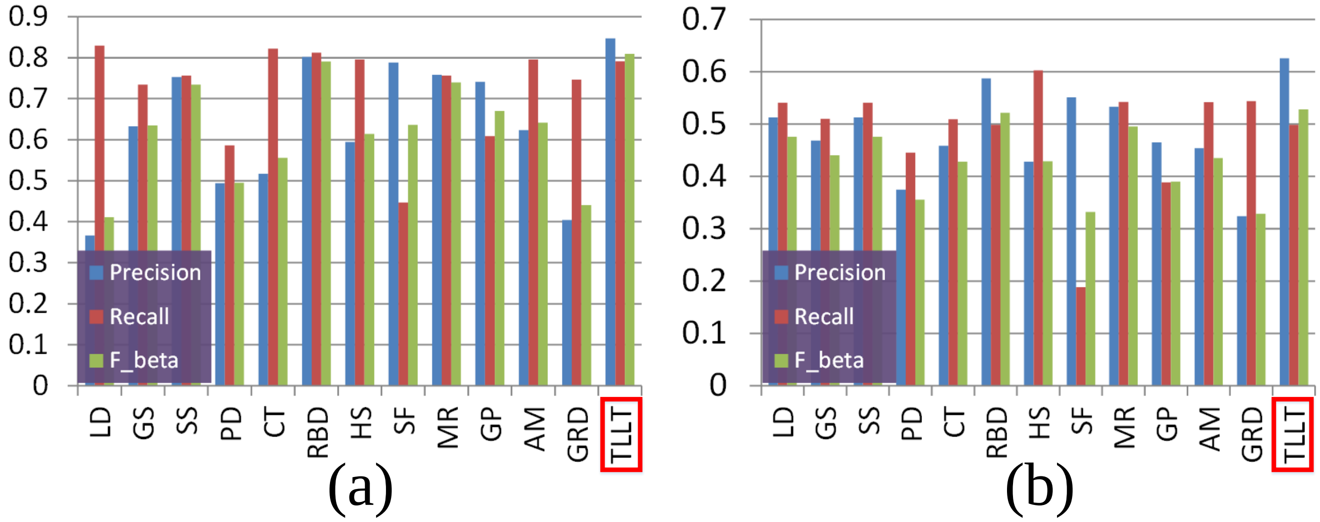

Figure 9: Comparison of different methods on two saliency detection datasets. (a) is MSRA 1000, and (b) is ECSSD.

suitable curriculum in every iteration, it needs relatively longer computational time. The iteration times for a normal image under our parametric settings are usually 5∼15. However, better results can be obtained as shown in Fig. 9, at the cost of more computational time.

To further present the merits of the proposed approach, we provide the resulting saliency maps of evaluated methods on several very challenging images from the two datasets (see Fig. 8). Though the backgrounds in these images are highly complicated, or very similar to the foregrounds, TLLT is able to generate fairly confident and clean saliency maps. In other words, TLLT is not easily confused by the unstructured background, and can make a clear distinction between the complex background and the regions of interest.

5.3 Parametric Sensitivity

There are two free parameters in our algorithm to be manually tuned: Gaussian kernel width θ and the amount of superpixels N. We evaluate each of the parameters θ and N by examining F_β^w with the other one fixed. Fig. 10 reveals that F_β^w is not sensitive to the change of N, but heavily depends on the choice of θ. Specifically, it can be observed that the highest records are obtained when $\theta = 0.25$ on both datasets, so we adjust θ to 0.25 for all the experiments.

6 Conclusion

This paper proposed a novel approach for saliency propagation through leveraging a teaching-to-learn and learning-to-teach paradigm. Different from the existing methods that propagated the saliency information entirely depending on the relationships among adjacent image regions, the proposed approach manipulated the propagation sequence from simple regions to difficult regions, thus leading to more reliable propagations. Consequently, our approach can render a more confident saliency map with higher background suppression, yielding a better popping out of objects of interest. Our approach is inspired by the theoretical results in educational psychology, and can also be understood from the well-known physical diffusion laws. Future work may study accelerating the proposed method and meanwhile exploring more insightful learning-to-teach principles.

References

[1] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk. SLIC superpixels compared to state-of-the-art superpixel methods. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 34(11):2274–2282, 2012.

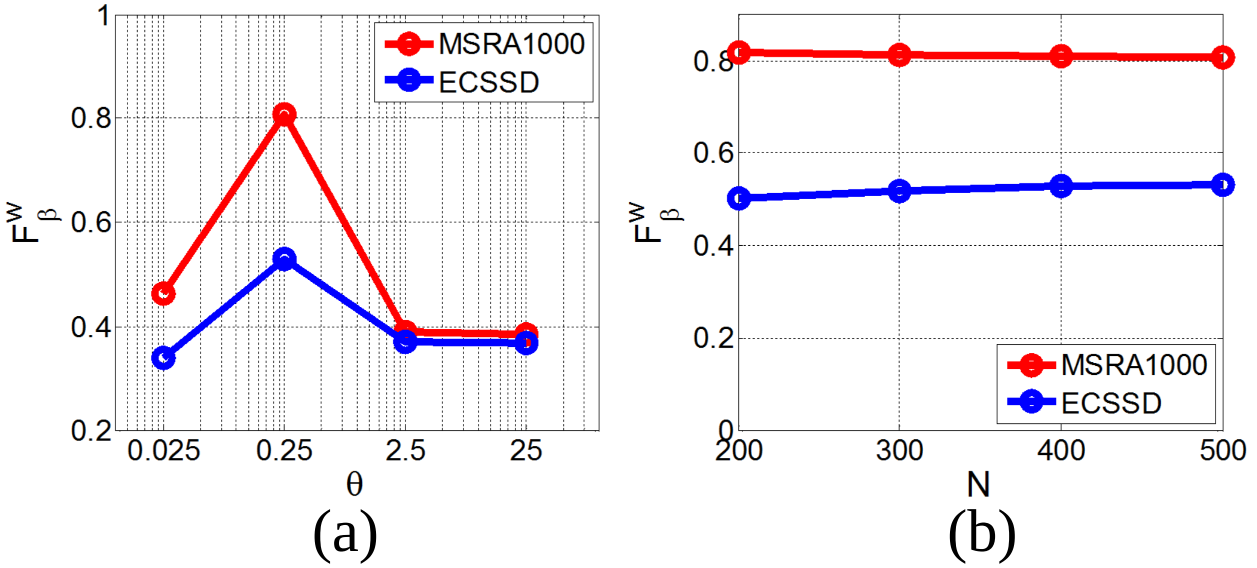

Figure 10: Parametric sensitivity analyses: (a) shows the variation of F_β^w w.r.t. θ by fixing $N = 400$; (b) presents the change of F_β^w w.r.t. N by keeping $\theta = 0.25$.

[2] Y. Bengio, J. Louradour, R. Collobert, and J. Weston. Curriculum learning. In *Proc. International Conference on Machine Learning*, pages 41–48. ACM, 2009.

[3] C. Bishop. *Pattern recognition and machine learning*, volume 1. springer New York, 2006.

[4] M. Cheng, G. Zhang, N. Mitra, X. Huang, and S. Hu. Global contrast based salient region detection. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 409–416. IEEE, 2011.

[5] J. Elman. Learning and development in neural networks: The importance of starting small. *Cognition*, 48(1):71–99, 1993.

[6] A. Fick. On liquid diffusion. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 10(63):30–39, 1855.

[7] K. Fu, C. Gong, I. Gu, and J. Yang. Geodesic saliency propagation for image salient region detection. In *Image Processing (ICIP), IEEE Conference on*, pages 3278–3282, 2013.

[8] K. Fu, C. Gong, I. Gu, J. Yang, and X. He. Spectral salient object detection. In *Multimedia and Expo (ICME), IEEE International Conference on*, 2014.

[9] V. Gopalakrishnan, Y. Hu, and D. Rajan. Random walks on graphs to model saliency in images. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1698–1705. IEEE, 2009.

[10] X. Hou and L. Zhang. Saliency detection: A spectral residual approach. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1–8. IEEE, 2007.

[11] L. Itti, C. Koch, and E. Niebur. A model of saliency-based visual attention for rapid scene analysis. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(11):1254–1259, 1998.

[12] B. Jiang, L. Zhang, H. Lu, C. Yang, and M. Yang. Saliency detection via absorbing markov chain. In *Computer Vision (ICCV), IEEE International Conference on*, pages 1665–1672. IEEE, 2013.

[13] F. Khan, B. Mutlu, and X. Zhu. How do humans teach: On curriculum learning and teaching dimension. In *Advances in Neural Information Processing Systems*, pages 1449–1457, 2011.

[14] J. Kim, D. Han, Y. Tai, and J. Kim. Salient region detection via high-dimensional color transform. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 883–890. IEEE, 2014.

[15] M. Kumar, B. Packer, and D. Koller. Self-paced learning for latent variable models. In *Advances in Neural Information Processing Systems*, pages 1189–1197, 2010.

[16] C. Lee, A. Varshney, and D. Jacobs. Mesh saliency. In *ACM Transactions on Graphics*, volume 24, pages 659–666. ACM, 2005.

[17] Y. Li, X. Hou, C. Koch, J. Rehg, and A. Yuille. The secrets of salient object segmentation. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 280–287. IEEE, 2014.

[18] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum. Learning to detect a salient object. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1–8. IEEE, 2007.

[19] R. Margolin, A. Tal, and L. Zelnik-Manor. What makes a patch distinct? In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1139–1146. IEEE, 2013.

[20] R. Margolin, L. Zelnik-Manor, and A. Tal. How to evaluate foreground maps. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 248–255. IEEE, 2014.

[21] S. Maybank. A probabilistic definition of salient regions for image matching. *Nuerocomputing*, 120(23):4–14, 2013.

[22] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung. Saliency filters: Contrast based filtering for salient region detection. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 733–740. IEEE, 2012.

[23] D. Rohde and D. Plaut. Language acquisition in the absence of explicit negative evidence: How important is starting small? *Cognition*, 72(1):67–109, 1999.

[24] Y. Wei, F. Wen, W. Zhu, and J. Sun. Geodesic saliency using background priors. In *European Conference on Computer Vision (ECCV)*, pages 29–42. Springer, 2012.

[25] W. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 1155–1162. IEEE, 2013.

[26] C. Yang, L. Zhang, and H. Lu. Graph-regularized saliency detection with convex-hull-based center prior. *Signal Processing Letters, IEEE*, 20(7):637–640, 2013.

[27] C. Yang, L. Zhang, H. Lu, X. Ruan, and M. Yang. Saliency detection via graph-based manifold ranking. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 3166–3173. IEEE, 2013.

[28] J. Yang and M. Yang. Top-down visual saliency via joint CRF and dictionary learning. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 2296–2303. IEEE, 2012.

[29] D. Zhou, J. Weston, A. Gretton, O. Bousquet, and B. Schölkopf. Ranking on data manifolds. *Advances in Neural Information Processing Systems*, 16:169–176, 2004.

[30] W. Zhu, S. Liang, Y. Wei, and J. Sun. Saliency optimization from robust background detection. In *Computer Vision and Pattern Recognition (CVPR), IEEE Conference on*, pages 2814–2821. IEEE, 2014.

[31] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *Proc. International Conference on Machine Learning*, volume 3, pages 912–919, 2003.