Salient Object Detection using Window Mask Transferring with Multi-layer Background Contrast

Quan Zhou[1], Shu Cai[1], Shaojun Zhu[2], and Baoyu Zheng[1]

[1]College of Telecom & Inf Eng, Nanjing Univ of Posts & Telecom, P.R. China
[2]Dept. of Comput & Inf Sci, University of Pennsylvania Philadelphia, PA, USA

Abstract. In this paper, we present a novel framework to incorporate bottom-up features and top-down guidance to identify salient objects based on two ideas. The first one automatically encodes object location prior to predict visual saliency without the requirement of center-biased assumption, while the second one estimates image saliency using contrast with respect to background regions. The proposed framework consists of the following three basic steps: In the top-down process, we create a specific location saliency map (SLSM), which can be identified by a set of overlapping windows likely to cover salient objects. The binary segmentation masks of training windows are treated as high-level knowledge to be transferred to the test image windows, which may share visual similarity with training windows. In the bottom-up process, a multi-layer segmentation framework is employed, which is able to provide vast robust background candidate regions specified by SLSM. Then the background contrast saliency map (BCSM) is computed based on low-level image stimuli features. SLSM and BCSM are finally integrated to a pixel-accurate saliency map. Extensive experiments show that our approach achieves the state-of-the-art results over MSRA 1000 and SED datasets.

1 Introduction

The human visual system (HVS) has an outstanding ability to quickly detect the most interesting regions in a given scene. In last few decades, the highly effective attention mechanisms of HVS have been extensively studied by researchers in the fields of physiology, psychology, neural systems, image processing, and computer vision [1–8], The computational modeling of HVS enables various vision applications, e.g., object detection/recognition [9, 10], image matching [2, 11], image segmentation [12], and video tracking [13].

Visual saliency can be viewed from different perspectives. Top-down (supervised) and bottom-up (unsupervised) are two typical categories. The first category often describes the saliency by the visual knowledge constructed from the training process, and then uses such knowledge for saliency detection on the test images [14, 15]. Based on the biological evidence that the human visual attention is often attracted to the image center [16], the center-biased assumption

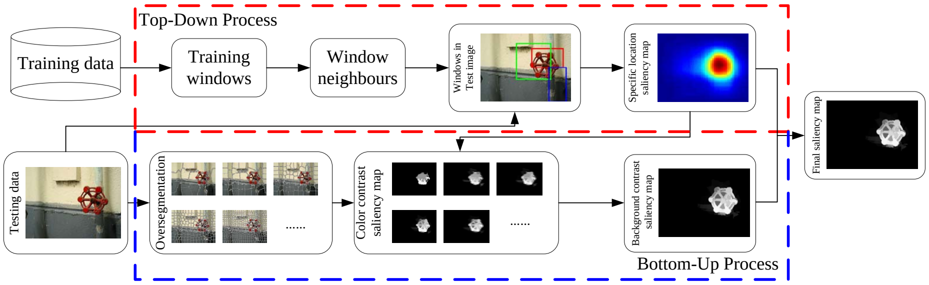

Fig. 1. Our approach consists of two components: (1) Top-down process. Given the training data consists of images with annotated binary segmentation masks, we first employ the technique of [9] to detect windows likely to contain salient objects on all training images and testing images. Then the binary segmentation masks of training windows are transferred to each detective windows in testing image with the most similar appearance (window neighbours). The transferred segmentation masks are used to derive the specific location saliency map (SLSM); (2) Bottom-up process. Using the over-segmentation technique of [25], an input testing image is first partitioned to multi-layer segmentation in a coarse to fine manner. Given the SLSM as prior map, a set of robust background regions are abstracted, and then the color-based contrast saliency maps are created for each layer of segmentation. These saliency maps are combined to form our background contrast saliency map (BCSM). SLSM and BCSM are finally integrated to a pixel-accurate saliency map. (Best viewed in color)

is often employed as the location prior for estimating visual saliency in top-down models [17, 15].

While the salient regions are mostly located in the image center, the inverse might not necessarily be true [18, 19]. Not all image center regions tend to be more salient. The salient object might be located far away from image center, even on the image boundary. Furthermore, a center-biased assumption always supposes that there is only one salient object within each image, yet it often fails when nature image contains two or more salient objects [19]. Thus, to detect salient regions without center-biased constrains, some semantic knowledge (e.g., face and pedestrian) are integrated in a top-down process, which is mostly based on object detectors [14, 20, 17]. The integration, however, acts rather more general on object category level than at the saliency-map level.

On the other hand, the bottom-up models are mainly motivated from the *contrast* formulation. For example, Itti *et al.* [1, 21] proposed a set of pre-attentive features including local center-surround intensity, color and direction contrasts. These contrasts were then integrated to compute image saliency through the winner-take-all competition. Cheng *et al.* [22] and Achanta *et al.* [23] utilize the global contrast with respect to the entire scene to estimate visual saliency. Recently, Borji and Itti [24] combine local and global patch rarities as contrast to measure saliency for eye-fixation task. We argue that the contrast based on background regions also plays an important role in such processes.

In this paper, we propose a novel method to integrate bottom-up, lower-level features and top-down, higher-level priors for salient object detection. Our approach is fully automatic and requires no center-biased assumption. The key idea of our top-down process is inspired by [26], where the binary segmentation masks are treated as prior information to be transferred from the supervised training image set to the testing image set. Then, the transferred segmentation masks are used to derive specific location prior of salient object in the test image.

Figure 1 illustrates the overview of our method. The basic intuition is that the windows with similar visual appearance often share similar binary segmentation masks. Since these transferred windows exhibit less visual variability than the whole scenes and are often centered on the salient regions, they are much suitable for location transfer with better support regions. As a result, we utilize the method of [9] to extract candidate windows that are likely to contain salient objects, and then transfer training window segmentation masks that share visual similarity to windows in the test image. Afterwards, the bottom-up saliency map is computed based on low-level image stimuli features. In nature images, although the salient regions and backgrounds may also tend to be perceptually heterogeneous, the appearance cues (e.g., color and texture) of the salient object region are still quite different from the backgrounds. Therefore, different from the previous methods that mainly utilize the local central-surround contrast [1, 15, 24] and global contrast [23, 22, 27] to encode saliency, our framework estimates visual saliency using the appearance-based contrast with respect to the background candidate regions. In order to automatically abstract robust background regions, we employ the multi-layer segmentation framework, which is able to provide large amount of background candidates within different sizes and scales.

The contributions of our approach are three-fold:

(1) In the top-down process, it proposes a specific location prior for salient object detection. Through window mask transferring, our method is able to provide more accurate location prior to detect salient regions, which results in more accurate and reliable saliency maps than the models using center-biased assumptions, such as [16] and [17];

(2) In the bottom-up process, unlike the previous approaches that utilize the local and global contrast to predict visual saliency, we attempt to estimate visual saliency using the contrast with respect to the background regions;

(3) Compared with most competitive models [1, 22, 14, 28, 17, 23, 18, 29, 27, 30, 31], our method achieves the state-of-the-art results over MSRA 1000 and SED datasets.

2 Related Work

In this section, we focus on reviewing the existing work for salient object detection, which can be roughly classified into two categories: bottom-up and top-down models.

The bottom-up approaches select the unique or rare subsets in an image as the salient regions [1, 32, 28, 33]. As a pioneer work, Itti *et al.* [1] introduced a biologically inspired saliency model based on the center-surround operation. Graph-based models [29, 34] are suggested to predict saliency following the principle of Markov random walk theory. Some researchers attempt to detect irregularities as visual saliency in the frequency domain [27, 23, 35]. Bruce and Tsotsos [36] established a bottom-up model following the principle of maximizing information sampled from a scene. Sparsity models [37, 17] are also employed to encode saliency, where the salient regions are identified as sparse noises when recovering the low-rank matrix.

Despite the success of these models, they are difficult to generalize to realword scenes. Instead, some researchers attempt to incorporate the top-down priors for salient object detection [15, 10, 20]. Li *et al.* [38] and Ma *et al.* [39] formulate the top-down factors as high level semantic cues (e.g., faces and pedestrian). Alternatively, Navalpakkam and Itti [40] modeled the top-down gain optimization as maximizing the signal-to-noise ratio (SNR). Liu *et al.* [15] proposed to adopt a conditional random field (CRF) model for predicting visual saliency. Bayesian modeling is also used for combining sensory evidence with prior constrains. In these models, the prior knowledge (such as scene context [14] or gist descriptors [41]) and sensory evidence (such as target features [42]), are probabilistically combined according to Bayesian rule. Different from these methods, our method employs the specific location prior as top-down knowledge by transferring window segmentation masks.

3 Our Approach

In this section, we elaborate on the details of our method. We first introduce how to obtain the specific location saliency map (SLSM) by transferring window masks. Given the multi-layer segmentations and SLSM on hand, we select a series of background regions that are used to compute background contrast saliency map (BCSM). Finally, two maps are incorporated to generate pixelwised saliency.

3.1 Specific Location Saliency Map (SLSM)

Finding Similar Windows. In order to utilize the prior knowledge of annotated binary segmentation mask in the training set, we first detect windows likely to contain an object using the "objectness" technique of [9]. It tends to return more windows covering an object with a well-defined boundary, rather than amorphous background elements. In our experiments, sampling only $\mathcal{N}$ windows per image (e.g., $\mathcal{N} = 100$) seems enough to cover most salient objects. Putting all the training windows together, we obtain the training window set $\{W_t\}$. This leads to retrieving much better neighborhood windows with similar appearance, whose segmentation masks are more suitable to transfer for test image. Given a new test image I as illustrated in Figure 2(a), the $\mathcal{N}$ number of "objectness" windows are also extracted using [9] as well as for the training images. Figure

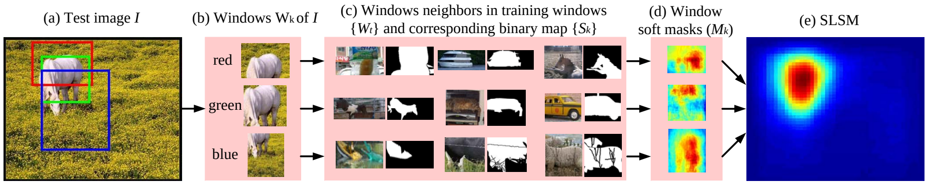
(a) Test image I
(b) Windows Wk of I
(c) Windows neighbors in training windows {Wt} and corresponding binary map {Sk}
(d) Window soft masks (Mk)
(e) SLSM
red
green
blue

Fig. 2. An example of the full pipeline for producing SLSM. Given a test image I in (a), three top windows (denoted as red, green and blue rectangles) are highlighted out of $\mathcal{N}$ windows, as shown in (b). The window neighbors are displayed in (c). It is shown that green window is tightly centered on an object and gets very good neighbors, while for red and blue windows, the neighbors are good matches for transferring segmentation mask, even though these windows do not cover the "horse" perfectly. This results in an accurate transfer mask for each window of I, as illustrated in (d). On the rightmost column of (e), we integrate the soft mask M_k from all windows into a soft mask for the whole scene, which is used to derive the SLSM. Note blue color denotes low saliency, while red color represents high saliency (Best viewed in color)

2(b) shows top three "objectness" windows in the test image I, and it is observed that many detective windows are centered on the salient object "horse". For one specific test window $W_k, k = \{1, 2, \cdots, \mathcal{N}\}$, we compute GIST feature [43] inside W_k to describe its appearance, and compare GIST descriptors with the ℓ^2-norm distance to all training windows $\{W_t\}$ to find window neighbors. Thus, the set $\{S_k^j\}, j = \{1, 2, \cdots, \mathcal{M}\}$ containing the segmentation masks of the top $\mathcal{M}$ training windows most similar to W_k is used for transferring. Figure 2(c) illustrates that the nearest neighbor windows accurately depict similar animals in similar poses, resulting in well-matched binary segmentation masks.

Segmentation Transfer. Let $S_T(x, y)$ be the SLSM, which defines the probability of pixel at location (x, y) to be salient. We construct $S_T(x, y)$ for each pixel from the segmentation masks transferred from all windows containing it.

1) Soft masks for windows. For the k^{th} test window W_k, we have a set of binary segmentation masks $\{S_k^j\}$ of neighbor windows from the training set. Here we compute a soft segmentation mask M_k for each W_k as the pixel-wise mean of the masks in $\{S_k^j\}$. To this end, all masks in $\{S_k^j\}$ are resized to the resolution of W_k. Let $\{S_k^{j'}\}$ be the resized masks, then the soft mask M_k for window W_k is defined as:

$$M_k = \frac{1}{\mathcal{M}} \sum_{j'=1}^{\mathcal{M}} S_k^{j'} \tag{1}$$

In this aligned space, a pixel value in M_k corresponds to the probability for it to be a salient object in $\{S_k^{j'}\}$. Figure 2(d) shows the corresponding M_k for the detected windows. Note the resolution of each soft window mask M_k is the same as the one of detected window in Figure 2(b).

2) Soft mask for the test image. After obtaining soft masks M_k, we integrate M_k for all windows into a single soft segmentation mask $M(x, y)$ for the test image I. For each window W_k, we place its soft mask M_k at the image location

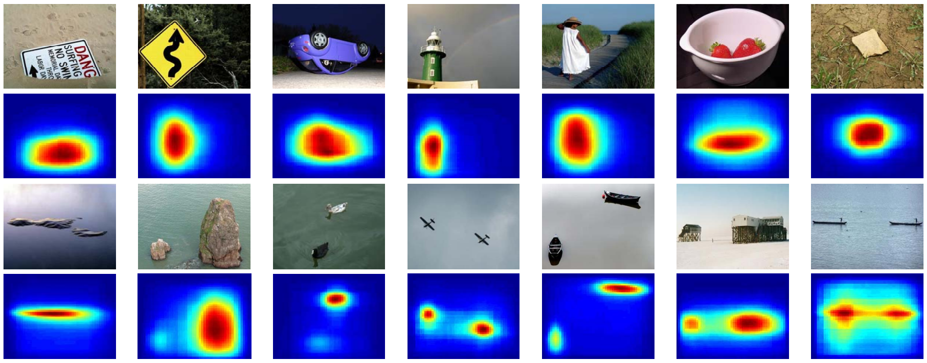

Fig. 3. Illustration of SLSM. The first and third rows are the example images form MSRA 1000 and SED dataset, respectively. The second and fourth rows are the corresponding SLSM, where blue color denotes low saliency, while red color represents high saliency. (Best viewed in color)

(x, y) defined by W_k. The soft mask $M(x, y)$ of the test image is the pixel-wise mean of all $\mathcal{N}$ placed masks $M_k(x, y)$:

$$M(x, y) = \frac{1}{\mathcal{N}} \sum_{k=1}^{\mathcal{N}} M_k(x, y) \tag{2}$$

A pixel value in $M(x, y)$ is the probability for it to be salient, according to all transferred segmentations (as illustrated in Figure 2(d)). Therefore, we define the SLSM $S_T(x, y)$ as

$$S_T(x, y) = M(x, y) \tag{3}$$

Due to the integration of all soft segmentation masks $M_k(x, y)$ from the individual windows, our approach achieves even more robust results. The key step of our approach is that we extract many windows (e.g., 100 per image) overlapping salient object. One effect is that a certain window might not have good neighbors in the training set, leading to transferring an inaccurate or even completely incorrect mask $M_k(x, y)$. However, other overlapping windows will probably have good neighbors, diminishing the effect of the inaccurate $M_k(x, y)$ in the integration step. Another effect may happen when the transferred windows may not cover a salient object, (e.g., detecting a patch on the backgrounds, as the blue window shown in Figure 1). This does not pose a problem to our approach, as the training images are decomposed in the same type of windows [9]. Therefore, a background window will probably also has similar appearance neighbors on backgrounds in the training images, resulting in correctly transferring a background binary segmentation mask. As a result, our approach is fully symmetric over salient and background windows. Figure 3 exhibits some SLSMs of nature images over MSRA 1000 and SED datasets.

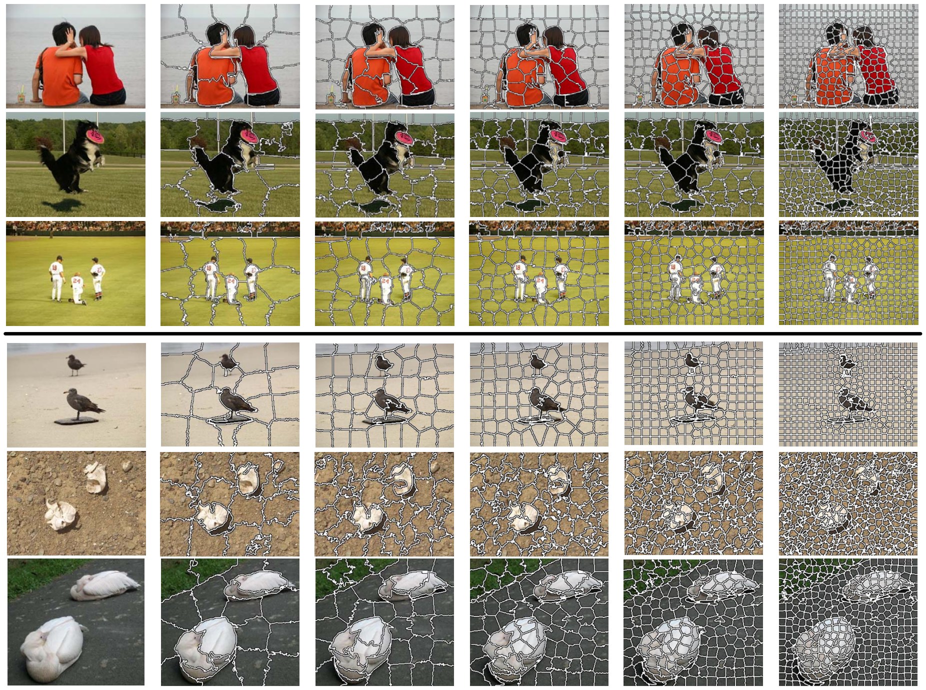

Fig. 4. Image representation by multi-layer segmentation. The upper panel shows the examples from MSRA dataset, while the bottom panel illustrates the examples from SED dataset. From left to right are the original images and their over-segmentation results in a coarse to fine manner. Different segments are separated by white boundaries.

3.2 Background Contrast Saliency Map (BCSM)

No matter where the salient object locates, it often exhibits quite different appearance cues (e.g., color and texture) within the entire scene. We thus build our background contrast saliency map (BCSM) guided by the global color-based contrast measurement [22]. Instead of computing saliency based on an entire image, here we calculate the contrast based on background candidates.

Multi-layer Segmentation. In order to make full use of background candidate regions, we employ the multi-layer segmentation framework.

Traditionally, an image is typically represented by a two-dimensional array of RGB pixels. With no prior knowledge of how to group these pixels, we can compute only local cues, such as pixel colors, intensities or responses to convolution with bank of filters [44, 45]. Alternatively, we use the SLIC algorithm [25] to implement over-segmentation, since it performs more efficiently. In practice, we partition test image I into J layers of segmentations. There are two parameters to be tuned for this segmentation algorithm, namely (rgnSize, regularizer), which denote the number of segments used for over-segmentation and the tradeoff appearance for spatial regularity, respectively. As shown in Figure 4, the advantages of using this technique are that it can often group the homogeneous regions with similar appearance and preserve the true boundary of objects.

Computing BCSM. Denote $S_B(x,y)$ as the BCSM to convey a sense of the dissimilarity of a pixel based on its local feature with respect to backgrounds, so that $S_B(x,y)$ also gives the probability of pixel at location (x,y) to be salient. We construct $S_B(x,y)$ for each pixel from the multi-layer segmentation via all segments containing it.

1) Color contrast saliency for each layer segmentation. Let r_i^j be i^{th} specific segment in j^{th} layer of segmentation. According to the SLSM, we select the segments with low saliency value to be background candidates, which are ready to compute the color-based contrast saliency. Let $\mathcal{B}^j = \{B_1^j, B_2^j, \cdots, B_M^j\}$ be selected background candidate regions in j^{th} layer segmentation. To measure how distinct the salient region is with respect to $B_m^j \in \mathcal{B}^j$, we can measure the distance between r_i^j and B_m^j using various visual cues such as intensity, color, and texture/texton. In this paper, we use the inverse cosine distance between histograms of HSV space to compute the color-based contrast:

$$C_{i,m}^j(\mathcal{H}(r_i^j), \mathcal{H}(B_m^j)) = 1 - \frac{\mathcal{H}(r_i^j)^T \mathcal{H}(B_m^j)}{||\mathcal{H}(r_i^j)|| \cdot ||\mathcal{H}(B_m^j)||} \tag{4}$$

where $\mathcal{H}(\cdot)$ is the binned histogram calculated from all color channels of one segment, and $||\cdot||$ denotes the ℓ^2 norm. We use histograms because they are a robust global description of appearance. They are insensitive to small changes in size, shape, and viewpoint. From Equation (4), it is observed that the contrast between r_i^j and B_m^j is very low when they look similar, otherwise not. For the given segment r_i^j, its color contrast saliency $S_B(r_i^j)$ is computed as the mean of L smallest contrasts in $\{C_{i,m}^j(\cdot,\cdot)\}, m = 1, 2, \cdots, M$

$$S_B(r_i^j) = \frac{1}{L} \sum_{m=1}^{L} C_{i,m}^j(\cdot,\cdot) \tag{5}$$

As will be seen, when r_i^j is truly a salient region, the L smallest contrasts always get large value with respect to the background regions, resulting in high saliency for $S_B(r_i^j)$. The saliency map $S_B(r_i^j)$ is normalized to a fixed range $[0, 255]$, and $S_B(r_i^j)$ is assigned to each image pixel belonging to r_i^j with the saliency value as $S_B^j(x,y)$.

2) BCSM for testing image. We now incorporate the $S_B^j(x,y)$ for all segmentation layers into a single saliency map for the test image I. Then the BCSM $S_B(x,y)$ is defined as:

$$S_B(x,y) = \frac{1}{J} \sum_{j=1}^{J} S_B^j(x,y) \tag{6}$$

$S_B(x,y)$ is also normalized to a fixed range $[0, 255]$.

3.3 Combined Saliency

We integrate SLSM and BCSM to produce our final saliency map $S(x, y)$ with a linearly combination model

$$S(x, y) = \eta \cdot S_T(x, y) + (1 - \eta) \cdot S_B(x, y) \tag{7}$$

where η is the harmonic parameter to balance the top-down SLSM and bottom-up BCSM. Then $S(x, y)$ is normalized to a fixed range $[0, 255]$.

4 Experimental Results

To validate our proposed method, we carried out several experiments on two benchmark datasets using the Precision-Recall curve and F-measure described below. The main reason behind employing several datasets is that current datasets have different image and feature statistics, stimulus varieties, and center-biases. Hence, it is necessary to employ several datasets as models leverage different features that their distribution varies across datasets.

Datasets. We test our proposed model on two datasets: (1) Microsoft Research Asian (MSRA) 1000 dataset [23] is the most widely used and as baseline benchmark for evaluating salient object detection models. It contains 1000 images with resolution of approximate 400×300 or 300×400 pixels, which only have one salient object per image and provides accurate object-contour-based ground truth. (2) The SED [19] dataset is a smaller dataset only containing 100 images with resolution ranged from 300×196 to 225×300 pixels. The reason to employ this dataset lies in that it is not center-biased and there are two salient objects in each image. Therefore, this dataset is more challenging for the task of salient object detection.

Baselines. To show the advantages of our method, we selected 12 state-of-the-art models as baselines for comparison, which are spectral residual saliency (SR [27]), spatiotemporal cues (LC [31]), visual attention measure (IT [1]), graph-based saliency (GB [29]), frequency-tuned saliency (FT [23]), salient region detection (AC [30]), context-aware saliency (CA [14]), global-contrast saliency (HC and RC [22]), saliency filter (SF [28]), low rank matrix recovery (LR [17]), and geodesic saliency (SP [18]). In practice, we implemented all the 12 state-of-the-art models using a Dual Core 2.6 GHz machine with 4GB memory over two datasets to generate saliency maps.

Evaluation Metrics. In order to quantitatively evaluate the effectiveness of our method, we conducted experiments based on the following widely used criteria. The precision-recall curve (PRC) is used to evaluate the similarity between the predicted saliency maps and the ground truth. Precision corresponds to the percentage of salient pixels correctly assigned, while recall corresponds to the fraction of detected salient pixels in relation to the ground truth number of salient pixels. Another criterion to evaluate the overall performance is the F-measure [23, 22], which is used to weight harmonic mean measurement of precision and

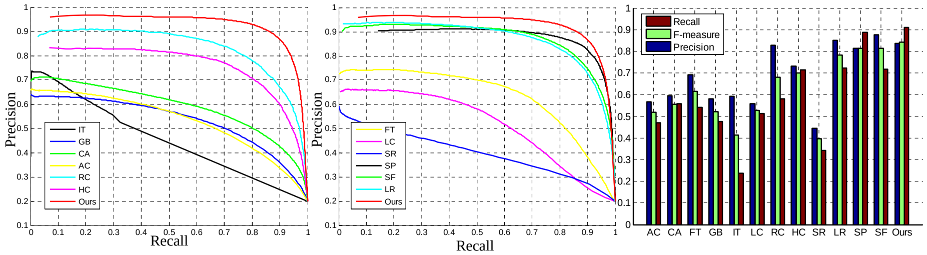

Fig. 5. Quantitative comparison for all algorithms with naive thresholding of saliency maps using 1000 publicly available benchmark images. Left and middle: PRC of our method compared with CA [14], AC [30], IT [1], LC [31], SR [27], GB [29], SF [28], LR [17], FT [23], SP [18], HC and RC [22]. Right: Average precision, recall and F-measure with adaptive-thresholding segmentation. Our method shows high precision, recall, and F_β values on the MSRA 1000 dataset. (Best viewed in color)

recall. The F-measure is defined as

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{8}$$

where $\beta^2 = 0.3$ following [23, 22].

Implemented Details. In order to make our approach work well, it is important to establish a large and diverse training set, which is able to provide sufficient appearance statistics to distinguish salient regions and background. In our implementation, we employ the full PASCAL VOC 2006 [46] as our training dataset, since it includes more than 5000 images with accurate annotated object-contour-based ground truth.

The parameter settings are: $\mathcal{N} = 100$ for sampling windows per image, $\mathcal{M} = 100$ for window neighbors for one sampling window, $J = 5$ for segmentation layers in a coarse to fine manner, $L = 5$ for computing color contrast saliency involved in Equation (5), (rgnSize, regularizer) are initialized as $\{25, 10\}$, and rgnSize is updated as $\{25, 50, 100, 200, 400\}$ with fixed regularizer, $\eta = 0.6$ to balance SLSM and BCSM for producing final saliency map.

We follow two widely used methodologies [23, 17] to implement our experiments. In the first implementation, we adopt the scheme that segments image according to the saliency values with a fixed threshold. Given a threshold $T \in [0, 255]$, the regions whose saliency values are higher than threshold are marked as a salient object. The segmented image is then compared with the ground truth to obtain the precision and recall. We draw the PRC using a series of precision-recall pairs by varying T from 0 to 255.

In the second implementation, the test image is segmented by an adaptive threshold method [17]. Given the over-segmented image, an average saliency is calculated for each segment. Then an overall mean saliency value over the entire image is obtained as well. If the saliency in this segment is larger than twice of the overall mean saliency value, the segment is marked as foreground. Precision

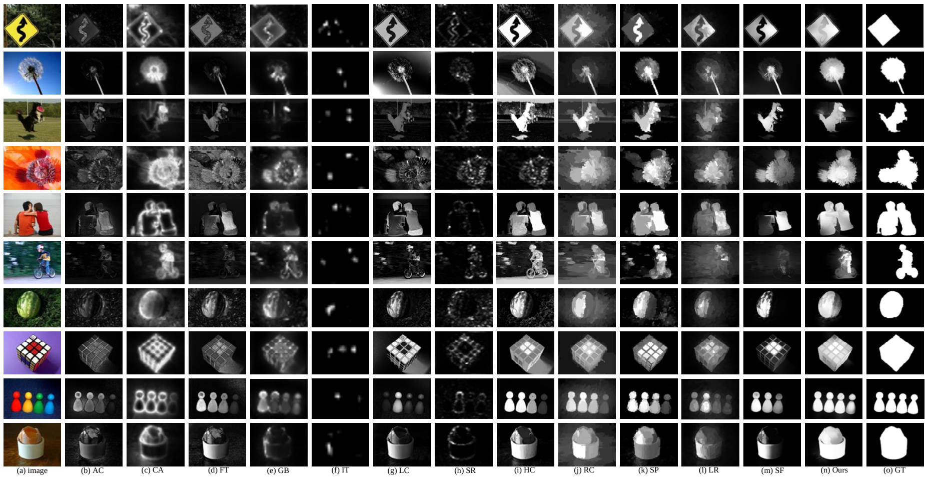

Fig. 6. Visual comparison of previous approaches with our method. See the legend of Figure 5 for the references to all methods.

and recall values are sequentially calculated, and F-measure is finally computed for evaluation.

Overall Results. The average PRC and F-measure on MSRA 1000 dataset are illustrated in Figure 5. It clearly shows that our method outperforms other approaches. It is interesting to note that the minimum recall value of our methods starts from 0.08, and the corresponding precision is higher than those of the other methods, probably because the saliency maps computed by our methods contain more pixels with the saliency value 255. The improvement of recall over other methods is more significant, which means our method are likely to detect more salient regions, while keeping a high accuracy.

We also evaluate our method on SED dataset and compare it with other 12 models. Figure 7(a) reports the comparison results in terms of F-measure. Our method achieves the state-of-the-art results and higher F-measure value (ours = 0.763) than other competitive models (SF = 0.739, LR = 0.68, RC = 0.62, and HC = 0.60), which clearly shows the validity of our approach in the case of more than one salient object within each image.

Visual comparison with different methods on MSRA 1000 dataset are shown in Figure 6, and some qualitative results on SED dataset are displayed in Figure 8. Compared with other models, our method is very effective in eliminating the cluttered backgrounds, and uniformly highlighted salient regions with well-defined object shapes, no matter whether salient objects locate in image center, or far away from image center, even on the image boundary.

Analysis of Implemental Efficiency. In order to evaluate the implemental efficiency of our method, we compare the average running time with some competitive models, and report the results in Table 1. Our method is slower than HC and FT, and faster than SR, IT, GB, SP, SF, RC, LR. The majority of this

Table 1. Average running time of different methods on MSRA 1000 dataset.

Method	SR [27]	IT [1]	GB [29]	FT [23]	SP [18]
Time(s)	0.064	0.611	1.614	0.016	1.213
Code	Matlab	Matlab	Matlab	C++	Matlab
Method	HC [22]	RC [22]	SF [28]	LR [17]	ours
Time(s)	0.019	0.253	0.153	1.748	0.759
Code	C++	C++	C++	Matlab	Matlab

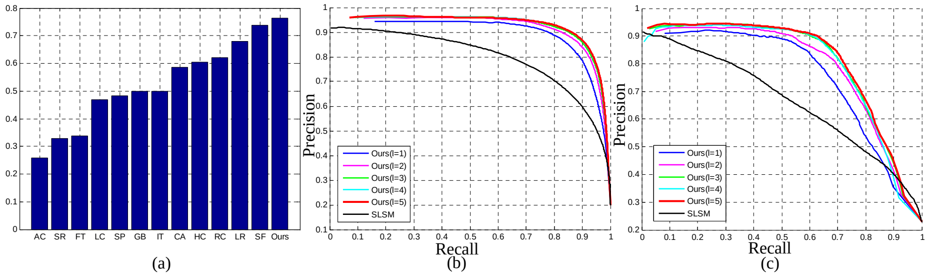

Fig. 7. Left: From left to right: (a) F-measure of the different saliency models to ground truth on SED dataset. (b) and (c) The comparison of PRC by gradually increasing the layers of segmentation on MSRA 1000 and SED dataset, respectively. In (b) and (c), the performance of individual SLSM is also included. (Best viewed in color)

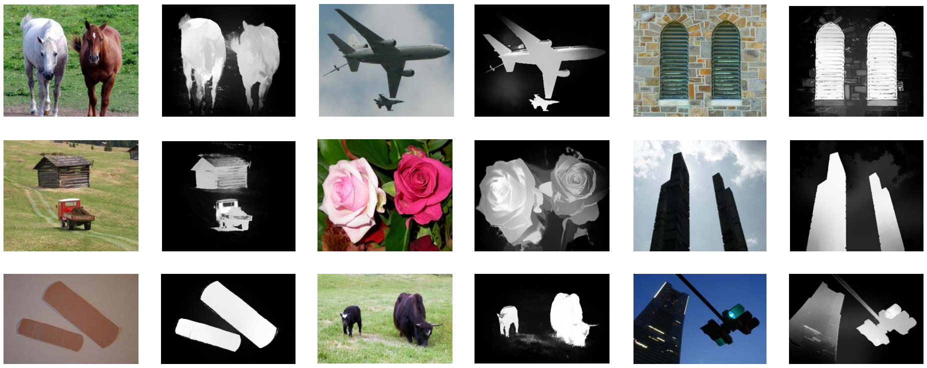

Fig. 8. Some visual examples on SED dataset. The first, third and fifth columns are original images, and the second, fourth and sixth columns are the corresponding saliency maps. (Best viewed in color)

time is spent performing multi-layer segmentation, producing detected window, and computing window neighbours (about 80%), and only 20% account for the actual saliency computation.

Analysis of Parameter Setting and Individual Saliency Map. One factor affecting the performance is the layers of over-segmentation. Figure 7(b) and (c) exhibit the plot of PRC with different number of segmentation layers on the MSRA 1000 and SED datasets, respectively. It is observed that better performance can be achieved along with the increasement of segmentation layers, and no further improvement after 5 layer segmentations. This demonstrates that our method performs robustly over a wide range of segmentation layers.

In order to evaluate the contributions of each individual saliency map, the PRC of SLSM is also included for comparison in Figure 7(b) and (c). The performance of SLSM is already better than most of the competitive models, whose results are shown in Figure 5. Using BCSM noticeably improves the performance for both two datasets, which indicates the importance of measuring visual saliency using contrast with respect to background regions.

5 Conclusion and Future Work

In this paper, we propose a novel framework for salient object detection based on two key ideas: (1) using specific location information as top-down prior by transferring segmentation masks from windows in the training images that are visually similar to windows in the test image; (2) using the contrast based on the background candidates in bottom-up process makes our method more robust than methods estimating contrast on the entire image. Compared with existing competitive models, the extensive experiments show that our approach achieves the state-of-the-art results over MSRA 1000 and SED datasets. In the future, we would like to combine two saliency maps, SLSM and BCSM, with adaptive weights using learning technique, as well as [47] does.

Acknowledgement. The authors would like to thank all the anonymous reviewers valuable comments. We would like to thank Prof. Liang Zhou for his valuable comments to improve the readability of the whole paper. This work was supported by NSFC 61201165, 61271240, 61401228, 61403350, PAPD and NY213067.

References

1. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. TPAMI **20** (1998) 1254–1259
2. Zhao, R., Ouyang, W., Wang, X.: Person re-identification by salience matching. In: ICCV. (2013) 73–80
3. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S.: Salient object detection: A discriminative regional feature integration approach. In: CVPR. (2013) 2083–2090
4. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: CVPR. (2013) 1155–1162
5. Borji, A., Tavakoli, H.R., Sihite, D.N., Itti, L.: Analysis of scores, datasets, and models in visual saliency prediction. In: CVPR. (2013) 921–928
6. Li, X., Lu, H., Zhang, L., Ruan, X., Yang, M.H.: Saliency detection via dense and sparse reconstruction, ICCV (2013)

7. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: CVPR. (2013) 3166–3173
8. Marchesotti, L., Cifarelli, C., Csurka, G.: A framework for visual saliency detection with applications to image thumbnailing. In: ICCV. (2009) 2232–2239
9. Alexe, B., Deselaers, T., Ferrari, V.: What is an object? In: CVPR. (2010) 73–80
10. Gao, D., Han, S., Vasconcelos, N.: Discriminant saliency, the detection of suspicious coincidences, and applications to visual recognition. TPAMI **31** (2009) 989–1005
11. Toshev, A., Shi, J., Daniilidis, K.: Image matching via saliency region correspondences. In: CVPR. (2007) 1–8
12. Jung, C., Kim, C.: A unified spectral-domain approach for saliency detection and its application to automatic object segmentation. TIP **21** (2012) 1272–1283
13. Mahadevan, V., Vasconcelos, N.: Saliency-based discriminant tracking. In: CVPR. (2009) 1007–1013
14. Goferman, S., Zelnik-Manor, L., Tal, A.: Context-aware saliency detection. In: CVPR. (2010) 2376–2383
15. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.: Learning to detect a salient object. TPAMI **33** (2011) 353–367
16. Tatler, B.: The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. Journal of Vision **7** (2007) 1–17
17. Shen, X., Wu, Y.: A unified approach to salient object detection via low rank matrix recovery. In: CVPR. (2012) 853–860
18. Wei, Y., Wen, F., Zhu, W., Sun, J.: Geodesic saliency using background priors. In: ECCV. (2012)
19. Borji, A., Sihite, D.N., Itti, L.: Salient object detection: A benchmark. In: ECCV. (2012) 414–429
20. Judd, T., Ehinger, K., Durand, F., Torralba, A.: Learning to predict where humans look. In: ICCV. (2009) 2106–2113
21. Itti, L., Koch, C.: A saliency-based search mechanism for overt and covert shifts of visual attention. Vision research **40** (2000) 1489–1506
22. Cheng, M., Zhang, G., Mitra, N., Huang, X., Hu, S.: Global contrast based salient region detection. In: CVPR. (2011) 409–416
23. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: CVPR. (2009) 1597–1604
24. Borji, A., Itti, L.: Exploiting local and global patch rarities for saliency detection. In: CVPR. (2012) 478–485
25. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Süsstrunk, S.: Slic superpixels. EPEL, Tech. Rep **149300** (2010)
26. Kuettel, D., Ferrari, V.: Figure-ground segmentation by transferring window masks. In: CVPR. (2012) 558–565
27. Hou, X., Zhang, L.: Saliency detection: A spectral residual approach. In: CVPR. (2007) 1–8
28. Perazzi, F., Krahenbuhl, P., Pritch, Y., Hornung, A.: Saliency filters: Contrast based filtering for salient region detection. In: CVPR. (2012) 733–740
29. J., H., C., K., P., P.: Graph-based visual saliency. In: NIPS. (2006) 545–552
30. Achanta, R., Estrada, F., Wils, P., Susstrunk, S.: Salient region detection and segmentation. Computer Vision Systems (2008) 66–75
31. Zhai, Y., Shah, M.: Visual attention detection in video sequences using spatiotemporal cues. In: ACMMM. (2006) 815–824
32. Parkhurst, D., Law, K., Niebur, E.: Modeling the role of salience in the allocation of overt visual attention. Vision research **42** (2002) 107–124

33. Wang, W., Wang, Y., Huang, Q., Gao, W.: Measuring visual saliency by site entropy rate. In: CVPR. (2010) 2368–2375
34. Gopalakrishnan, V., Hu, Y., Rajan, D.: Random walks on graphs for salient object detection in images. TIP **19** (2010) 3232–3242
35. Guo, C., Ma, Q., Zhang, L.: Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In: CVPR. (2008) 1–8
36. Bruce, N., Tsotsos, J.: Saliency based on information maximization. In: NIPS. (2006) 155–162
37. Lang, C., Liu, G., Yu, J., Yan, S.: Saliency detection by multi-task sparsity pursuit. TIP **21** (2012) 1327–1338
38. Li, J., Tian, Y., Huang, T., Gao, W.: Probabilistic multi-task learning for visual saliency estimation in video. IJCV **90** (2010) 150–165
39. Ma, Y.F., Hua, X.S., Lu, L., Zhang, H.J.: A generic framework of user attention model and its application in video summarization. TMM **7** (2005) 907–919
40. Navalpakkam, V., Itti, L.: Search goal tunes visual features optimally. Neuron **53** (2007) 605–617
41. Torralba, A., Oliva, A., Castelhano, M., Henderson, J.: Contextual guidance of eye movements and attention in real-world scenes: the role of global features in object search. Psychological review **113** (2006) 766–786
42. Zhang, L., Tong, M., Marks, T., Shan, H., Cottrell, G.: Sun: A bayesian framework for saliency using natural statistics. Journal of Vision **8** (2008) 1–20
43. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV **42** (2001) 145–175
44. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. TPAMI **24** (2002) 603–619
45. Shi, J., Malik, J.: Normalized cuts and image segmentation. TPAMI **22** (2000) 888–905
46. Everingham, M., Zisserman, A., Williams, C.K.I., Van Gool, L.: The PASCAL Visual Object Classes Challenge 2006 (VOC2006) Results. (http://www.pascal-network.org/challenges/VOC/voc2006/results.pdf)
47. Itti, L., Koch, C.: Feature combination strategies for saliency-based visual attention systems. Journal of Electronic Imaging **10** (2001) 161–169