Multimedia Tools and Applications (2020) 79:10873–10897
https://doi.org/10.1007/s11042-019-07970-x

Exploiting color name space for salient object detection

Jing Lou[1] · Huan Wang[2] · Longtao Chen[2] · Fenglei Xu[2] · Qingyuan Xia[2] · Wei Zhu[2] · Mingwu Ren[2]

Received: 19 December 2018 / Revised: 8 June 2019 / Accepted: 10 July 2019 /
Published online: 26 December 2019

Abstract
In this paper, we will investigate the contribution of color names for the task of salient object detection. An input image is first converted to color name space, which is consisted of 11 probabilistic channels. By exploiting a surroundedness cue, we obtain a saliency map through a linear combination of a set of sequential attention maps. To overcome the limitation of only using the surroundedness cue, two global cues with respect to color names are invoked to guide the computation of a weighted saliency map. Finally, we integrate the above two saliency maps into a unified framework to generate the final result. In addition, an improved post-processing procedure is introduced to effectively suppress image backgrounds while uniformly highlight salient objects. Experimental results show that the proposed model produces more accurate saliency maps and performs well against twenty-one saliency models in terms of three evaluation metrics on three public data sets.

Keywords Saliency · Salient object detection · Figure-ground segregation · Surroundedness · Color names · Color name space

1 Introduction

Visual attention, one of intrinsic properties of human vision to extract important information from abundant visual inputs, is concerned with the understanding and modeling of biological perception systems. Psychophysical and physiological studies indicate that the selective

✉ Jing Lou
loujing@jsut.edu.cn
http://www.loujing.com/cns-sod/

✉ Mingwu Ren
renmingwu@mail.njust.edu.cn

[1] School of Information Engineering, Changzhou Vocational Institute of Mechatronic Technology, Changzhou, Jiangsu 213164, People's Republic of China

[2] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, People's Republic of China

attention mechanism, which can be directed by human visual system to gaze the most conspicuous location and then shift to the next conspicuous location, plays an important role in the early representation [20]. Since these conspicuous locations might be the feature cues based salient regions, the computational visual attention aims to deal with the automatic saliency detection in images or videos. In computer vision, the main tasks of saliency research include eye fixation prediction which attempts to predict human fixation data [11, 17, 18, 23, 31, 52], and salient object detection for the localization and identification of salient regions in visual scenes [3, 7, 48–50].

Over the past decades, saliency detection has been widely used in many computer vision applications, including image segmentation [34], object detection [26], object recognition [36], visual tracking [5], image and video compression [15], and video summarization [21]. Generally, the resultant map of saliency detection is called "saliency map", which topographically describes the conspicuity of each location in the whole scene. From a computational point of view, saliency detection techniques can be divided into two categories: slow, top-down, task-dependent manner; and rapid, bottom-up, task-independent manner [32]. Although top-down manner is indispensable for guiding the attention to behaviorally relevant objects, the salient features based bottom-up attention is more closely related to an early stage of visual processing [20, 44] and has been investigated by numerous researchers.

In the feature integration theory of attention, a visual scene is initially coded along a number of elementary features, e.g., color, orientation, brightness, and spatial frequency [44]. The selective attention mechanism [20] suggests to compute these elementary features in parallel and combine the resultant cortical topographic maps into a saliency map. Hence, a majority of bottom-up saliency models aim to investigate different visual features and apply them to define the saliency of a pixel or a region. In these models, the contrast based detection is one of the most commonly adopted techniques. As no prior knowledge regarding salient objects is provided, the contrast based detection mainly focuses on two aspects: local center-surround difference, and global rarity.

For local center-surround difference, one of the most influential bottom-up saliency models is introduced by Itti et al. [18]. Basing on the Koch and Ullman's early representation model [20], Itti et al. extract various features at multiple resolutions and use center-surround differences between different resolutions to form a saliency map. Ma and Zhang [27] regard an image as a perceive field and define the saliency by measuring differences between the stimuli perceived by different perception units. Goferman et al. [12] exploit four basic principles of human visual attention to detect the context-aware saliency, i.e., local low-level features, global considerations, visual organization rules, and high-level factors. Furthermore, by means of the Kullback-Leibler divergence, an information-theoretic approach is proposed to extract saliency from multi-scale center-surround feature distributions [19].

For another, the global rarity based saliency models tend to find rare features from an image. Achanta et al. [3] propose a frequency-tuned (FT) approach, which defines the pixel-wise saliency by comparing the color of each pixel with the average image color in LAB color space. In [7], Cheng et al. present a histogram contrast (HC) based saliency method, which uses color statistics to compute saliency. In addition, a regional contrast (RC) based saliency method is introduced in that work, which simultaneously evaluates global contrast differences and spatial coherences. In order to reduce the complexity of calculating the color contrasts between regions, we subsequently follow the RC method and propose a regional principal color (RPC) based saliency method [25] by only retaining the most frequently occurred color of each superpixel. Besides the widely used color features, some other visual

cues are also exploited in the global contrast based saliency models, such as orientation [43], intensity [51], spectrum [17, 23], and texture [38].

In this paper, we also focus on the bottom-up and contrast-based saliency detection technique. Actually, if we review the task of salient object detection, we can see it has two clear implications: one is that the detected regions should be salient in an image, the other is that these salient regions should contain objects of any category. Gestalt psychological studies indicate that objects lying in the foreground may result in being more salient than background elements [30, 37]. Since salient objects are more likely to be involved in foreground regions, two questions consequently arise: 1) How to extract foreground regions? 2) How to define the contrast-based saliency? For the first question, one answer is to employ figure-ground segregation.

Recently, a simple and effective saliency model called "Boolean Map based Saliency" (BMS) is proposed in [52]. The BMS model first demonstrates that the rarity based models sometimes ignore global structure information and falsely highlight high contrast regions. Then following the suggestion of Gestalt psychology that the surroundedness may influence figure-surround segregation [33], BMS exploits a set of randomly sampled boolean maps to model the saliency of foreground regions. By using different parameter settings, BMS is suitable for both eye fixation prediction and salient object detection, and achieves the state-of-the-art performance.

Here, we only discuss its results of salient object detection. Although three channels of LAB color space are chosen as the randomly sampled feature maps, the essence of BMS is the use of the closed outer contours of foreground regions. The effect of salient object detection is somewhat equivalent to applying it to a lightness image. As illustrated in Fig. 1, it is interesting that if we convert the input RGB image (Fig. 1a) to LAB color space and apply BMS to the L channel (normalized to [0, 255], see Fig. 1d), we obtain two similar saliency maps (cf. Fig. 1b and e). The detected salient regions have similar characteristics, that is, they are enclosed by the outer boundaries and not connected to the image borders. Obviously, the color information is discarded in this case.

In this paper, we couple a surroundedness cue with two global color cues into a unified framework by extending BMS to *Color Name Space*, which is obtained by using the PLSA-bg color naming model [46] (or called PLSA-ind in [47]). In computer vision, color

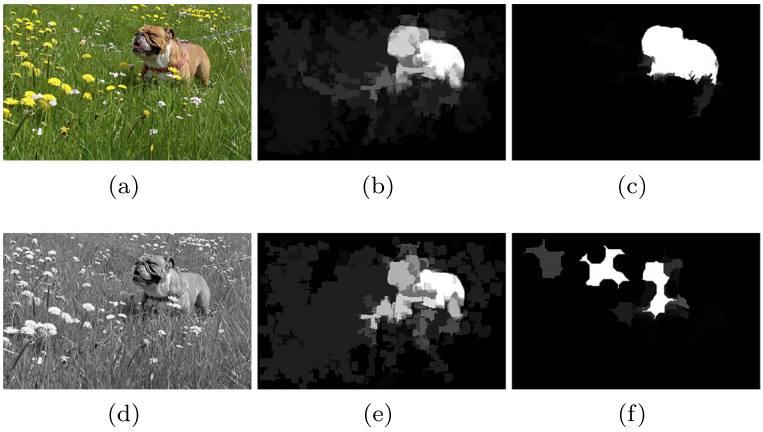


Fig. 1 **a** RGB image from ECSSD data set [40, 48], and the saliency maps generated by using **b** BMS [52] and **c** our model, respectively. **d** The L channel obtained by converting **a** to LAB color space, and the resultant saliency maps of **e** BMS and **f** our model

names are linguistic color labels assigned to image pixels. The linguistic study of Berlin and Kay [4] indicates that there are eleven basic color terms (i.e., color names) in the English language, as given in Table 1. In the proposed model, both the probabilities and statistics of eleven color names are simultaneously incorporated to measure color differences. The topological structure information also participates in the computation of the color names based saliency, hence several weighted master attention maps are generated. Through a simple linear combination and an improved post-processing procedure, we obtain two saliency maps and then fuse them into a single map. Finally, several image processing procedures, including truncation operation, intensity mapping, and hole-filling, are invoked to infer the final result. Figure 1c and f show the saliency results produced by the proposed model. We can see that the color name space based saliency shows higher precision. It demonstrates that the color cue is of as much importance as the surroundedness cue.

In the following sections, the proposed model will be called "CNS". The main contributions of this paper include:

1) By expoliting color name space, we propose an integrated framework to effectively compute the color based saliency.
2) A weighted global contrast mechanism is introduced to incorporate more color cues into the topological structure information of an image.
3) An improved post-processing procedure is proposed to uniformly highlight salient objects, which are easy to be segmented.

The remainder of this paper is organized as follows. Section 2 is the review of related work. The proposed salient object detection model is presented in Section 3. In Section 4, performance comparisons are made with three benchmark data sets. Conclusions and possible extensions are presented in Section 5.

2 Related work

We base the proposed model on BMS [52] and PLSA-bg [46]. The key idea of BMS is the use of the surroundedness cue, which can be characterized by a set of boolean maps. The BMS model first converts an input RGB image I to LAB color space, then scales each channel to [0, 255]. Subsequently, BMS chooses each channel as a feature map, and uses a set of fixed thresholds to binarize each feature map to boolean maps B_i as follows [52]:

$$B_i = \mathbf{THRESH}(\phi(I), \theta) \ , \tag{1}$$

where $\phi(I)$ is a feature map of I, and θ represents a fixed threshold. Based on a Gestalt principle of figure-ground segregation [33], BMS performs several morphological operations

Table 1 Eleven basic color terms in the English language

i	1	2	3	4	5
Term (t_i)	black	blue	brown	grey	green
RGB (c_i)	[0 0 0]	[0 0 1]	[.5 .4 .25]	[.5 .5 .5]	[0 1 0]
6	7	8	9	10	11
Orange	pink	purple	red	white	yellow
[1 .8 0]	[1 .5 1]	[1 0 1]	[1 0 0]	[1 1 1]	[1 1 0]

to generate a set of attention maps, in which all the regions connected to the image borders are masked out since they are not surrounded by any closed outer contour. The final saliency map is simply the average of these attention maps, followed by a morphological post-processing.

The surroundedness cue is also invoked in the proposed CNS model. However, different from BMS, CNS uses color name space instead of LAB color space. In the field of document analysis, the standard PLSA model [16] computes the conditional probability of a word w in a document d, and estimates the distributions $p(z|d)$ and $p(w|z)$ by using an Expectation-Maximization (EM) algorithm, where z represents a latent topic. Considering that PLSA does not exploit the color name labels of training images, the PLSA-bg model [46] represents an image d (i.e., document) as a LAB color histogram with a group of color bins (i.e., words), and decomposes d into the foreground distribution according to a given color name label l_d (i.e., topic) and the background distribution shared between all training images. By estimating the mixing proportion of foreground versus background, color name distributions, and background model, the probability of a color name for a given image pixel is represented as

$$p(z|w) \propto p(z)p(w|z)\,, \tag{2}$$

where the prior over eleven color names is taken to be uniform.

Besides the probability information of color names, the proposed model makes use of a statistical cue. This is achieved by a *Color Name Histogram*, in which eleven color name bins are involved for measuring color differences. In [7], the HC method directly uses color statistics to define the saliency value of each color bin. Compared with HC, our model solely exploits the color name histogram to compute eleven weighting coefficients, and further produces eleven weighted master attention maps. The color name histogram does not participate in the generation of original attention maps, which are still determined by the surroundedness cue as used in BMS.

3 Color name space based saliency detection

To incorporate more color information, we extend BMS [52] from LAB color space to color name space. Two saliency cues, i.e., surroundedness and color, are separately invoked to produce two saliency maps. They are fused into a single map for generating the final result. These steps are described in the following sections.

3.1 General framework

As illustrated in Fig. 2, the integrated framework of CNS includes two computational pipelines.

Pipeline I An input RGB image is first resized to 400 pixels in width and converted to color name space. The resultant space is composed of 11 monochrome intensity components, namely *Color Name Channel* in this paper. Following BMS [52], a set of attention maps is generated based on figure-ground segregation. The attention maps of each channel are linearly fused to produce a master attention map. Finally, the mean attention map $\bar{A}$ is obtained by combining 11 master attention maps and further post-processed to generate the saliency map S.

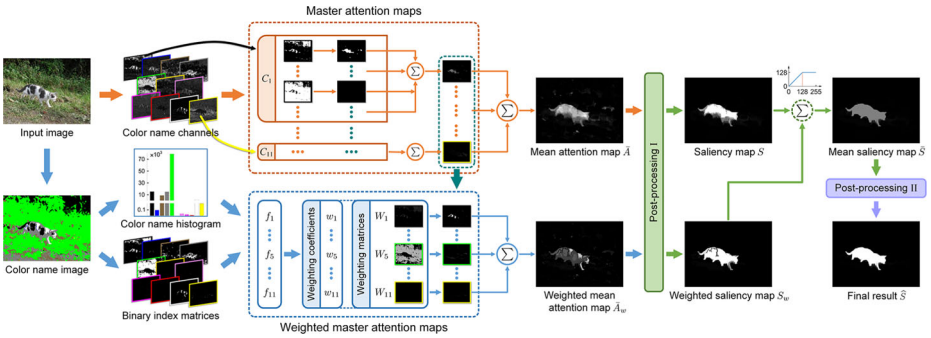


Fig. 2 Framework of the proposed CNS model

Pipeline II The resized RGB image is first converted to a *Color Name Image*, from which two statistical characteristics are derived: 1) a color name histogram which consists of 11 color bins, and 2) 11 binary index matrices where each of them represents the distribution of a corresponding color name. By exploiting two kinds of weighting patterns, we generate 11 weighted master attention maps. All the master attention maps obtained in Pipeline I also participate in this process. Finally, the weighted saliency map S_w is obtained by using the same combination and post-processing as used in Pipeline I.

Combination The saliency maps S and S_w are fed into a truncation operation to produce the mean saliency map $\bar{S}$, which simultaneously codes for the topological structure and color conspicuity over the entire image. In addition, we apply another post-processing procedure to generate the final saliency result $\widehat{S}$, in which the salient region is evenly highlighted and smoothed for convenience in the task of salient object segmentation.

3.2 Color name channel based attention map

First, we directly use the **im2c** function provided by [46][1] to generate the color name space $\mathbf{C} = \{C_1, C_2, \ldots, C_{11}\}$, where each color name channel C_i has the range of values [0, 1]. Thus, for the resized RGB image I, the color representation of each pixel is mapped from a 3-dimensional (3-D) RGB value to a probabilistic 11-dimensional (11-D) vector which sums up to 1. Considering that the topological structure of I is independent of the perceptual color coherence, each color name channel is treated equally and normalized to [0, 255] for the subsequent thresholding operation.

Then, we use a set of sequential thresholds from 0 to 255 with a step size of δ to binarize each color name channel $C_i \in \mathbf{C}$ to n boolean maps

$$B_i^j = \texttt{THRESH}\left(C_i,\ \theta_j\right), \tag{3}$$

where at each threshold θ_j, the above function generates a boolean map B_i^j by setting all the values above θ_j to 1s and replacing all the others with 0s. After two morphological operations on B_i^j, including closing and hole-filling, we use a clear-border algorithm [41] to mask out all the foreground regions connected to the image borders, and obtain a corresponding attention map A_i^j. The same processing steps are also executed for the complement map of

[1] http://lear.inrialpes.fr/people/vandeweijer/color_names.html

B_i^j (denoted $\widetilde{B}_i^j$). As summarized in Algorithm 1, two parameters are required in this stage: sample step δ, and kernel radius ω_c of the closing operation. We will discuss the influences of them in Section 4.3.

Algorithm 1 Attention map computation.

Input: resized RGB image I
Output: attention maps A_i^j and $\widetilde{A}_i^j$
convert I from RGB space to color name space $\mathbf{C}$
for each $C_i \in \mathbf{C}$ **do**
  for $\theta_j = 0 : \delta : 255$ **do**
    $B_i^j = \mathtt{THRESH}(C_i, \theta_j)$
    $B_i^j = \mathtt{CLOSE}(B_i^j, \omega_c)$
    $B_i^j = \mathtt{FILL}(B_i^j)$
    $A_i^j = \mathtt{CLEAR-BORDER}(B_i^j)$
    $\widetilde{B}_i^j = \mathtt{INVERT}(B_i^j)$
    $\widetilde{B}_i^j = \mathtt{CLOSE}(\widetilde{B}_i^j, \omega_c)$
    $\widetilde{B}_i^j = \mathtt{FILL}(\widetilde{B}_i^j)$
    $\widetilde{A}_i^j = \mathtt{CLEAR-BORDER}(\widetilde{B}_i^j)$
  end for
end for

However, different from BMS which averages all the attention maps, the proposed model computes the mean attention map A_i of each color name channel C_i separately. All the attention maps A_i^j and $\widetilde{A}_i^j$ share the same weight and are averaged to A_i, which is called *Master Attention Map* in this paper. Then, the mean attention map $\bar{A}$ of 11 master attention mpas can be further calculated as follows:

$$A_i = \frac{1}{2n} \sum_{j=1}^{n} \left(A_i^j + \widetilde{A}_i^j \right) , \tag{4}$$

$$\bar{A} = \frac{1}{11} \sum_{i=1}^{11} A_i \, . \tag{5}$$

Actually, if we merge (4) and (5), we can get the same computation procedure of $\bar{A}$ as introduced in the BMS model [52]. The slight difference lies in the usage of 11 master attention maps. In Pipeline I, the computation of $\bar{A}$ is mainly based on the surroundedness cue. To make better use of color name space, the proposed framework couples the surroundedness cue with two color cues to compute the color based saliency. In Section 3.4, we will again use the 11 master attention maps to produce a weighted mean attention map $\bar{A}_w$.

3.3 Post-processing

The mean attention map $\bar{A}$ is shown in Fig. 3a. Due to the existence of other surrounded objects that have clear boundaries and uniform colors (for example, the red flower below the cat), there are several small salient regions in $\bar{A}$. In order to outstand the main salient object (i.e., the cat), we also follow BMS to remove small salient regions by sequentially

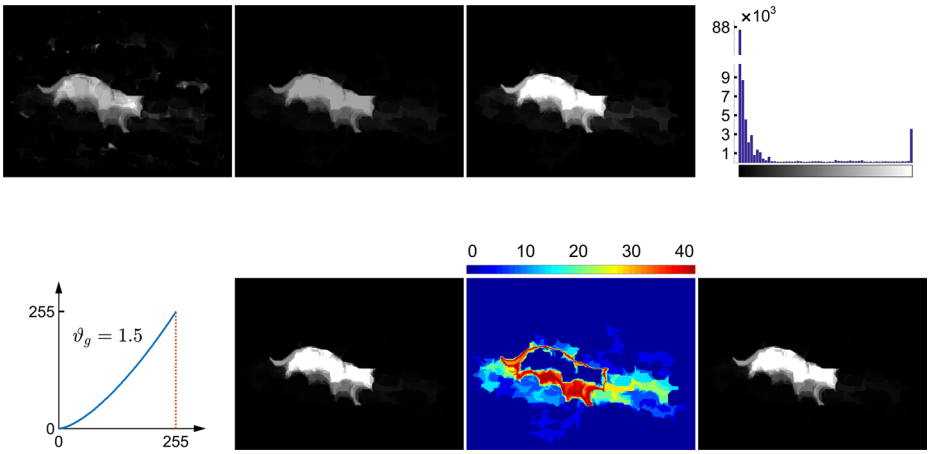


Fig. 3 Post-processing I. **a** Mean attention map $\bar{A}$. **b** Morphological reconstruction. **c** Normalization result, and **d** its histogram. **e** Intensity mapping curve. **f** Mapping result obtained by using $\vartheta_r = 0.02$ and $\vartheta_g = 1.5$. **g** Difference between **c** and **f**. **h** Saliency map S

performing two steps of morphological reconstruction operations [14, 45]. The structuring element used here is a disk shape with the radius ω_r. Figure 3b shows the reconstruction result. It can be observed that those small salient regions have been erased while the original shape of the salient cat is still remained.

For the task of salient object detection, the ideal output should be a binary map in which the pixel values of salient objects are 1s while the others are 0s. However, the disadvantage of the morphological reconstruction is that the high intensity values of salient pixels are suppressed simultaneously. In addition, the background of the reconstruction result contains some inconspicuous regions with non-black pixels, which would decrease the detection precision. To address the above issues, a nonlinear mapping function is introduced to transform the intensity values to a new range. Overall, we wish to weight the mapping toward the lower output values and map all the intensity values above a specific threshold to 1s. Suppose that F is the reconstruction result, the intensity mapping function has the syntax form as follows:

$$G = \texttt{MAP}\left(F,\ [0, T_F],\ \vartheta_g\right), \tag{6}$$

where T_F is the truncation threshold, ϑ_g determines the mapping relationship between F and G. To suppress non-salient pixels, the lower limit of the mapping is set to 0, and ϑ_g is set to be greater than 1.

In (6), all the intensity values above T_F (i.e., in the interval $[T_F, 255]$) are clipped and mapped to 1s. For automatically obtaining T_F, we exploit the statistical information extracted from the histogram of F. After scaling F to [0, 255] (see Fig. 3c), we get its histogram H where $H_k, k \in [0, 255]$ denotes the number of pixels at the kth intensity level. By summing up the number of pixels from H_0, we obtain the minimum intensity level T_F which should satisfy the following criteria:

$$(1 - \vartheta_r) \sum_{k=0}^{255} H_k \ \leqslant \ \sum_{k=0}^{T_F} H_k\,, \tag{7}$$

that is, the non-salient pixels should cover no less than $1 - \vartheta_r$ of the total number of image pixels. For convenience, we abbreviate (6) as

$$G = \mathbf{MAP}\left(F,\ \vartheta_r,\ \vartheta_g\right), \tag{8}$$

where ϑ_r is empirically set to be less than 10%.

Figure 3e illustrates the intensity mapping curve with $\vartheta_r = 0.02$ and $\vartheta_g = 1.5$. By using this mapping, the lower (darker) values in the output map (Fig. 3f) are further suppressed. From the difference map (Fig. 3g), we can see that those non-salient regions on the right side of the cat have been eliminated. Finally, we perform a hole-filling operation to generate the saliency map S. The whole post-processing procedure is summarized in Algorithm 2. In Section 4.3, we will discuss the influences of the parameters ω_r, ϑ_r, and ϑ_g.

Algorithm 2 Post-processing I.

Input: mean attention map $\bar{A}$
Output: saliency map S
1: $S = \mathbf{RECONSTRUCT}\,(\bar{A},\ \omega_r)$
2: $S = \mathbf{NORMALIZE}\,(S,\ [0, 255])$
3: $S = \mathbf{MAP}\,(S,\ \vartheta_r,\ \vartheta_g)$
4: $S = \mathbf{FILL}\,(S)$

3.4 Global color cue based saliency

As indicated previously, we then introduce a color-based saliency algorithm to overcome the limitation of only using the surroundedness cue. In order to take advantage of color attributes, two global color cues including statistic and contrast, are inferred from color name image and employed to compute weighting coefficients and matrices. The 11 master attention maps obtained in Section 3.2 are coupled with two kinds of weights to further produce a weighted saliency map S_w.

First, we again use the **im2c** function [46] to convert each pixel value of the resized RGB image I from a 3-D RGB value to a probabilistic 11-D vector. By exploiting the index number of the largest element in the vector, we construct an index map $\mathbf{M}$ where each pixel has an integer value between 1 and 11. Basing on $\mathbf{M}$, we derive two kinds of weights.

3.4.1 Color name statistic based weights

If we use the corresponding RGB value c_i given in Table 1 to represent each pixel in $\mathbf{M}$, we get the color name image as shown in Fig. 4a. The histogram of the color name image has totally 11 color levels, where the ith level corresponds to the number of pixels having the color name t_i. In this paper, the histogram is called *Color Name Histogram*, as shown in Fig. 4b. From the color name histogram, we can obtain 11 probability values. The probability of the ith color name is denoted as f_i.

Another cue is the distributions of the color names in $\mathbf{M}$. For the purpose of combining with 11 master attention maps, we use (9) to construct 11 index matrices. In the ith index matrix M_i, any element value equal to i is set to 1, otherwise is set to 0:

$$M_i(x, y) = \begin{cases} 1, & \text{if } \mathbf{M}(x, y) = i\,; \\ 0, & \text{otherwise}\,. \end{cases} \tag{9}$$

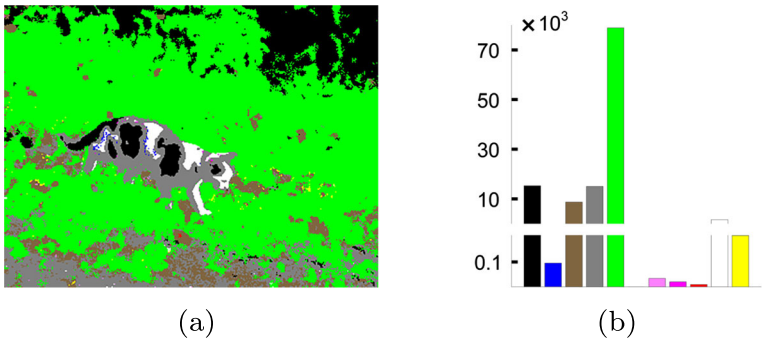


Fig. 4 Color name image and histogram. **a** Color name image. **b** Color name histogram

As discussed in Section 3.2, the attention map A_i of the ith color name channel is computed by linearly averaging boolean maps, where all the foreground regions that connected to the image borders are abandoned. For any boolean map, all the pixels in the surrounded regions share the same weight. To jointly consider the frequencies and distributions of different color names, we simply combine f_i and M_i to obtain the first kind of weights, i.e., weighting matrices

$$W_i = f_i M_i \,. \tag{10}$$

3.4.2 Color name contrast based weights

Mainly inspired by [51] and [7], we calculate the second kind of weights, i.e., the contrast based weighting coefficients. The weight of each color name is defined as its color contrast to all the other color names. All the pixels having the same color name share the same weight. For the color distance metric, we directly use the RGB values of 11 color names given in Table 1. Specifically, the weighting coefficient w_i of the color name t_i is defined as

$$w_i = \sum_{j=1}^{11} f_j \left\| c_i - c_j \right\|_2^2 , \tag{11}$$

where $\left\| c_i - c_j \right\|_2$ is the ℓ_2-norm of the color difference between the color names t_i and t_j.

By integrating two kinds of weights into 11 master attention maps and averaging them, we compute the weighted mean attention map $\bar{A}_w$ (see Fig. 5a) as follows:

$$\bar{A}_w = \sum_{i=1}^{11} w_i \cdot \mathbf{N}(W_i \circ A_i) \,, \tag{12}$$

where $\circ$ denotes the Hadamard product, and $\mathbf{N}(\cdot)$ is the normalization function which sets the values in $\bar{A}_w$ to [0, 1].

Figure 5b–e illustrate the same post-processing procedure introduced in Section 3.3. From Fig. 5h, we can see that the hole-filling operation completes the closed dark regions inside the cat. Finally, we obtain the second saliency map, i.e., the weighted saliency map S_w with the range [0, 255], as shown in Fig. 5g.

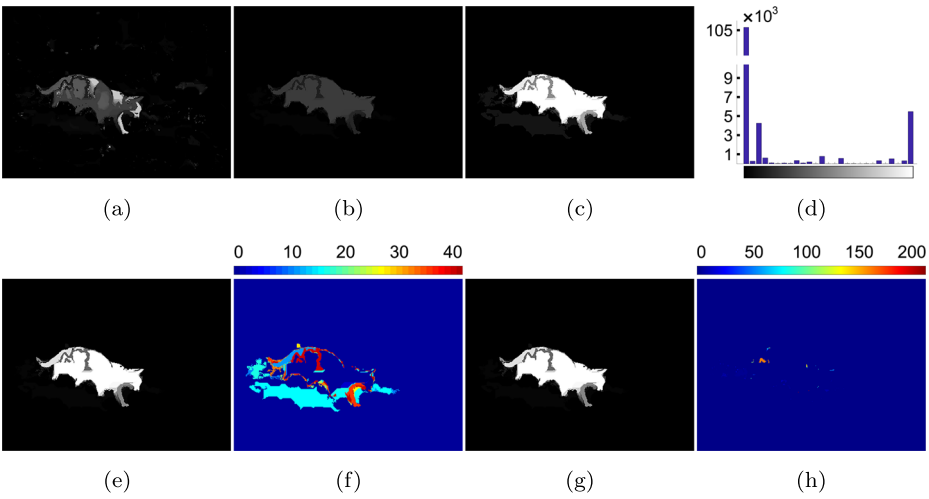


Fig. 5 Applying post-processing I to $\bar{A}_w$. **a** Weighted mean attention map $\bar{A}_w$. **b** Morphological reconstruction. **c** Normalization result, and **d** its histogram. **e** Intensity mapping result. **f** Difference between **c** and **e**. **g** Weighted saliency map S_w. **h** Difference between **e** and **g** for the demonstration of hole-filling

3.5 Combination

To couple with the saliency maps S and S_w, we simply average them at the first step of the combination stage. The original output is illustrated in Fig. 6a. Considering that the use of saliency maps is to assist in salient object segmentation, the original combination result is obviously not ideal. First, for the purpose of eliminating the perceptually insignificant regions outside the cat, we perform an intensity mapping in the post-processing I, which simultaneously suppresses the inner saliency and subsequently results in an indeterminate object region in S. Second, in S_w the salient object has a clear contour, but apparently shows a nonuniform intensity distribution. Third, the locations of the regions with higher saliency values are completely different between two saliency maps.

To address the above issues, a truncation operation is introduced to clip the original output. Intuitively, we wish the resultant salient object to have a uniform intensity distribution, which can be further highlighted by using a post-processing procedure. Since both S

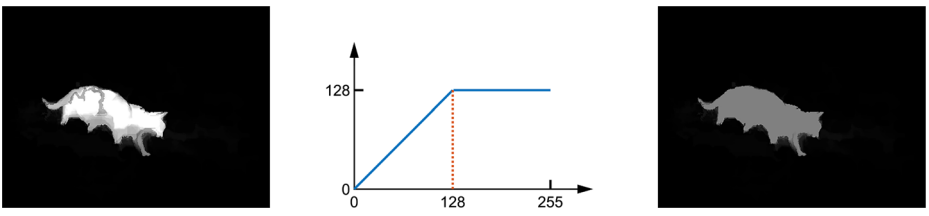


Fig. 6 Combination. **a** Original output of averaging Fig. 3h and g. **b** Truncation curve. **c** Mean saliency map $\bar{S}$

and S_w have been normalized to the range [0, 255], we define the improved mean saliency map $\bar{S}$ as

$$\bar{S} = \frac{[S + S_w]_0^{255}}{2}, \tag{13}$$

where $[\cdot]_0^{255}$ is the operator for truncating the inner to have values between 0 and 255.

As illustrated in Fig. 6b, the above definition causes a piecewise mapping, in which the values above 128 are clipped and the others stay unchanged. From Fig. 6c, we can see the resultant map $\bar{S}$ occupies the common salient parts between S and S_w. Although the detected region has lower saliency, the whole region is uniform and clearly stands out of the background. This means that we also can perform a post-processing operation on $\bar{S}$ to refine its saliency.

Algorithm 3 Post-processing II.

Input: mean saliency map $\bar{S}$
Output: final result $\widehat{S}$
1: $\widehat{S} = \texttt{MAP}(\bar{S}, \vartheta_r, \vartheta_g)$
2: $\widehat{S} = \texttt{FILL}(\widehat{S})$
3: $\widehat{S} = \texttt{RESIZE}(\widehat{S})$

A new post-processing procedure is summarized in Algorithm 3. Compared with Algorithm 2, the difference is that this procedure only includes two operations: intensity mapping and hole-filling. For the former operation, we use the same parameter settings as before. Figure 7a shows the histogram of the mean saliency map $\bar{S}$, the intensity mapping curve is illustrated in Fig. 7b. Note that different from Fig. 3e, here the intensity mapping curve maps the inputs in the range [0, 128] to the outputs in the range [0, 255]. After filling all the small dark holes in the object region, we obtain the final saliency result $\widehat{S}$ of the proposed model, as shown in Fig. 7c. It can be seen that our model well suppresses the image background and uniformly highlights the foreground object. More importantly, for the future task of salient object segmentation, we can easily perform a thresholding operation on $\widehat{S}$ while generate more stable segmentation results over a wide range of thresholds.

4 Experiments

We evaluate the proposed model with twenty-one saliency models including AC [2], BMS [52], CA [12], COV [11], FES [42], FT [3], GC [8], GU [8], HC [7], HFT [23],

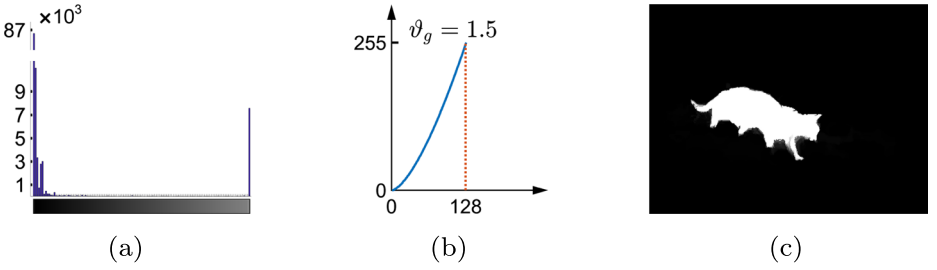


Fig. 7 Post-processing II. **a** Histogram of Fig. 6c. **b** Intensity mapping curve. **c** Final result $\widehat{S}$

MSS [1], PCA [28], RC [7], RPC [25], SEG [35], SeR [39], SIM [31], SR [17], SUN [53], SWD [10], and TLLT [13] on three benchmark data sets: ASD [3, 24], ECSSD [40, 48], and ImgSal [22, 23]. The used saliency maps of the above models are from:

- For BMS,[2] HFT,[3] HS,[4] RPC,[5] and TLLT[6] over all the data sets, we use the author-provided saliency maps, or run the authors' codes to obtain saliency maps.
- For the AC, CA, FT, HC, RC, and SR models on the ASD data set, we directly use the saliency maps provided by Cheng et al. [7].[7] For the remainder models on ASD, we retrieve the related saliency maps from the MSRA10K database [9].[8]
- For the remainder models, we employ the implementation of the salient object detection benchmark published by Borji et al. [6].[9] On the ECSSD data set, the saliency maps come directly from the author-provided results; on the ImgSal data set, we run the authors' source code to generate saliency maps.

4.1 Data sets

The popular ASD data set (a.k.a, MSRA1000) is a subset of MSRA5000 [24].[10] The original MSRA5000 data set contains 5000 images with the labeled rectangles from nine participants. Achanta et al. [3] consider that the use of saliency maps is for salient object segmentation, then derive ASD with 1000 images from MSRA5000. Instead of the user-drawn rectangles around salient regions used in [24], the ASD data set provides the object-contour based ground truth for more accurate comparisons of segmentation results.[11]

The ECSSD data set is an extension of CSSD.[12] In order to represent more general situations of natural images than ASD, Yan et al. construct the CSSD data set, which contains 200 images with diversified patterns in both foreground and background [48]. Subsequently, they extend CSSD to a larger data set named ECSSD, which includes 1000 structurally complex images and pixel-wise ground truth masks labeled by five helpers [40].

In addition, we evaluate the proposed model on the ImgSal data set, which is designed for the detection of salient regions of different sizes [22, 23]. The ImgSal contains 235 images collected using Google, and provides both region ground truth (human labeled) and fixation ground truth (by eye tracker). For region ground truth, the authors ask nineteen naive subjects to label the images in a random manner, and generate two kinds of labeling results for each image: binary map and probability map. In our experiments, we only use the binary maps for evaluating saliency detection results.

[2] http://cs-people.bu.edu/jmzhang/BMS/BMS.html
[3] http://www.escience.cn/people/jianli/DataBase.html
[4] http://www.cse.cuhk.edu.hk/leojia/projects/hsaliency/
[5] http://www.loujing.com/rpc-saliency/
[6] http://www.escience.cn/people/chengong/Codes.html
[7] http://cg.cs.tsinghua.edu.cn/people/~cmm/Saliency/Index.htm
[8] http://mmcheng.net/msra10k/
[9] http://mmcheng.net/salobjbenchmark/
[10] http://research.microsoft.com/en-us/um/people/jiansun/SalientObject/salient_object.htm
[11] http://ivrl.epfl.ch/supplementary_material/RK_CVPR09/
[12] http://www.cse.cuhk.edu.hk/leojia/projects/hsaliency/dataset.html

4.2 Experimental setup

The common used metrics to evaluate salient object detection models are *Precision-Recall* and F_β*-measure*. For an input image, the resultant saliency map is a gray-scale image having integer values in the range $[0, 255]$. So we can partition it to a binary map M by using a threshold, then compute precision and recall by comparing M with the corresponding ground truth G as follows:

$$Precision = \frac{|M \cap G|}{|M|}, \ Recall = \frac{|M \cap G|}{|G|}, \tag{14}$$

where $|\cdot|$ indicates the number of the foreground pixels. Moreover, to jointly evaluate precision and recall, the F_β-measure can be computed as

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \tag{15}$$

where β^2 is set to 0.3 for emphasizing the precision as suggested in [3].

In our experiments, two binarization ways are used to partition saliency maps.

1) Fixed Thresholding: We vary a threshold T_f from 0 to 255 to compute the scores of precision, recall and F_β-measure. Besides plotting the Precision-Recall and F_β-measure curves, we report two statistics for quantitative evaluation, i.e., average F_β score (denoted "AvgF") and maximum F_β score (denoted "MaxF").

2) Adaptive Thresholding: As presented in [3], we use an adaptive threshold T_a (cf. (16)) to partition $\widehat{S}$ and compute the scores of precision, recall and F_β-measure. Besides plotting Precision-Recall bars, we also report the F_β score obtained by using T_a (denoted "AdpF"):

$$T_a = \frac{2}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} \widehat{S}(x, y), \tag{16}$$

where W and H are the width and height of $\widehat{S}$ respectively, $\widehat{S}(x, y)$ is the saliency value at the coordinate (x, y).

4.3 Parameter analysis

The proposed model includes five parameters: sample step δ, kernel radius ω_c of closing operation, kernel radius ω_r of morphological reconstruction, saturation ratio ϑ_r and gamma ϑ_g of intensity mapping. To find the optimal parameter setting, we exploit the "MaxF" metric suggested in [29] to compare the saliency maps obtained using different parameter settings. After 256 F_β scores have been computed by fixed thresholding, the maximum one is chosen as the best score for each group of parameter setting. In our experiments, the ranges of five parameters are: $\delta \in [4 : 4 : 40]$, $\omega_c \in [1 : 20]$, $\omega_r \in [1 : 20]$, $\vartheta_r \in [0.001 : 0.001 : 0.009] \cup [0.01 : 0.01 : 0.1]$, and $\vartheta_g \in [1.0 : 0.1 : 3.0]$, respectively.

Figure 8 shows the influences of five parameters on the evaluation data sets. First, the proposed model is not sensitive to the parameter ϑ_g, while varying it from 1.0 to 3.0 rarely changes the MaxF scores. Second, the parameters ω_c, ω_r, and ϑ_r have direct impacts on MaxF, especially on the ImgSal data set. Overall, each MaxF curve shows a slight upward trend as the parameter value increases, then starts to drop after the MaxF reached the summit. Compared with ASD or ECSSD, the influences of the above three parameters are more apparent on the ImgSal data set. Third, the sample step δ does not significantly impact on MaxF, all the curves do not clearly show the unimodal distributions. However, the runtime

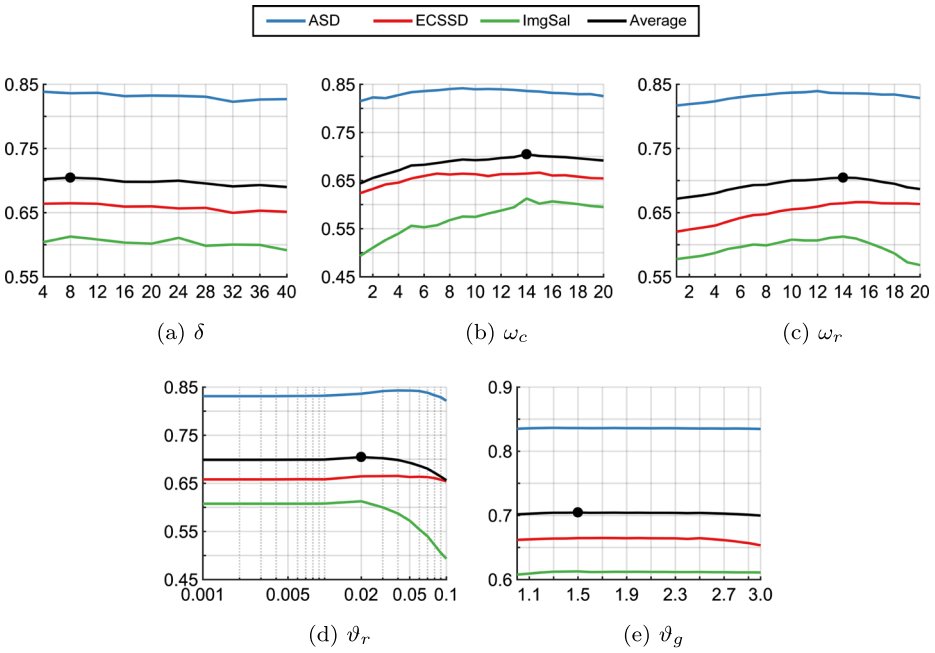


Fig. 8 Parameter analysis of the proposed model

of the proposed model is directly influenced by the sample step. As the value of δ decreases, it typically results in lower speed performance.

Based on the diversity of three data sets, we further use the average MaxF metric to determine the optimal parameter values. After three MaxF curves of each parameter have been obtained, we simply average them and choose the location of the maximum as the optimal value of each parameter. The black curves in Fig. 8 (indicated by "Average") exhibit the trends of five parameters. The optimal values of five parameters are reported in Table 2.

4.4 Results

We present the statistical comparison results of the proposed model compared with twenty-one saliency models. Figure 9a and b show the precision-recall and F_β-measure curves produced by fixed thresholding. The precision-recall bars generated by utilizing the adaptive threshold T_a are presented in Fig. 9c. More quantitative details are given in Fig. 11.

Table 2 Optimal parameter values

Parameter	Optimal value
δ	8
ω_c	14
ω_r	14
ϑ_r	0.02
ϑ_g	1.5

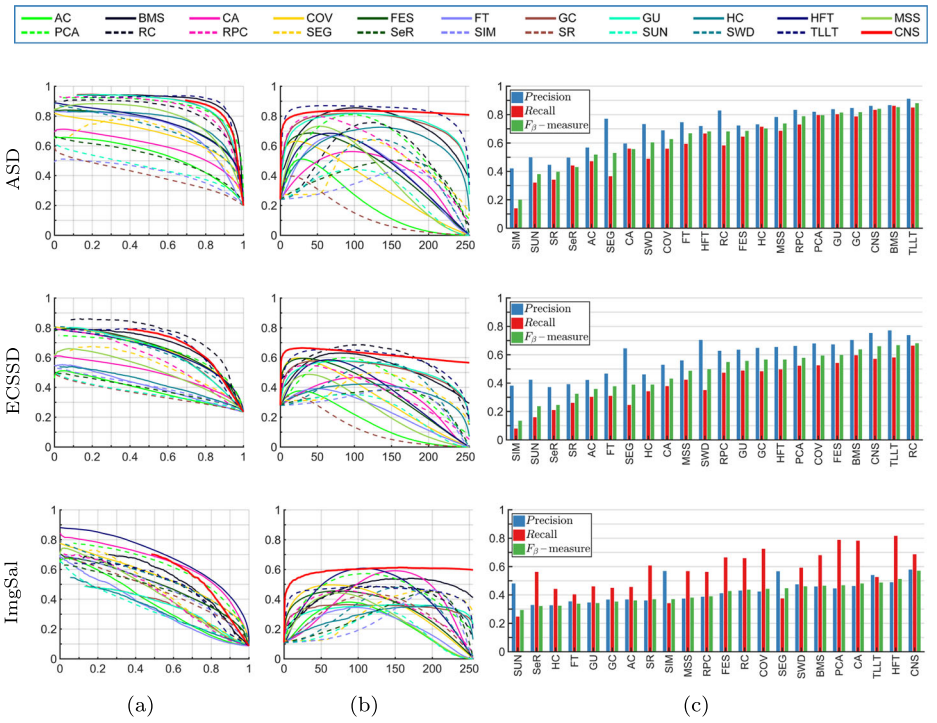


Fig. 9 Performance of the proposed model compared with twenty-one saliency models on ASD [3, 24], ECSSD [40, 48], and ImgSal [22, 23], respectively. **a** Precision-Recall curves. **b** F_β-measure curves. **c** Precision-Recall bars

Due to the intensity mapping used in the post-processing procedure, the resultant curves of our model clearly present two noticeable characteristics: one is that the recall scores span a more narrow range; the other is that the F_β-measure curves tend to be more flat after they rapidly reach the summits. Although having some disadvantages in the precision, our model has higher F_β scores, especially on the ECSSD and ImgSal data sets. The crucial advantage of our model indeed is associated with the essential task of salient object detection, which is to solve a salient foreground segmentation problem [6].

A good salient object detection model should generate accurate saliency maps with evenly highlighted foregrounds and thoroughly suppressed backgrounds. Then an easy way to extract salient objects is to binarize the saliency maps with a single fixed threshold. However, this threshold is quite difficult to determine automatically. In practice, we usually use the maximum F_β score (i.e., MaxF) to evaluate the performance of a saliency model, and choose the location of the MaxF as the optimal segmentation threshold [29]. Suppose that a saliency map is the same as its ground truth mask, the F_β-measure curve would be a horizontal line. Contrarily, if the F_β-measure curve is a horizontal line, we can obtain the identical segmentation results at any threshold in [0, 255]. Therefore, for two models having the same MaxF, we prefer to select the one which produces a more flat F_β-measure curve. This means that the segmentation results would be more stable (that is, virtually unchanged) over a wide range of thresholds.

Figure 10 shows a visual comparison of the saliency maps generated by different models. For these example images, our model generates more accurate saliency maps, which are very

close to the corresponding ground truth masks. The salient regions detected by our model have uniform intensities and well-defined boundaries, which result in a simple thresholding for the subsequent salient object segmentation.

In Fig. 11, we report the quantitative statistics of the three evaluation metrics discussed earlier. The baseline scores, indicated by "Average", are simply the average of evaluation scores. With respect to AvgF, TLLT ranks the first on ASD. The proposed model outperforms all the others on ECSSD and ImgSal. Obviously, this is mainly owed to more flat F_β-measure curves in a wide range.

However, on the ASD and ECSSD data sets, our model has some disadvantages in terms of MaxF and AdpF. For the MaxF metric, TLLT performs the best on the ASD data set, and ranks the third on ECSSD. For the AdpF metric, TLLT also ranks the first on ASD, while on ECSSD the RC model performs the best. However, our model is among top three models in terms of both MaxF and AdpF on these two data sets. On the ImgSal data set, our model again outperforms all the others with large margins. Moreover, compared with ASD and ECSSD, the average performances of all the models are lower on ImgSal. It means that this data set is more challenging because the images collected in it contain salient regions of different sizes.

Finally, on average, the proposed model performs the best over all the compared models. Besides, the best two models are TLLT and BMS. The MaxF scores of nine models are lower than the average score. The top five worst models are SUN, SR, AC, SIM, and SeR. Except AC, the other four are eye fixation prediction models, which have no advantages for salient object detection because the output saliency maps are blurred and sparse. But this does not necessarily mean that the eye fixation prediction models are not suitable for

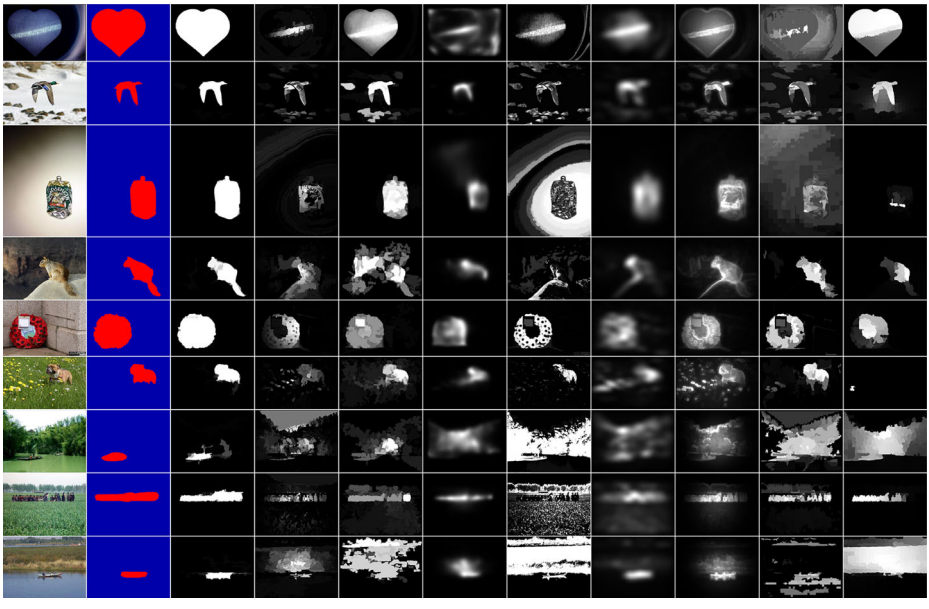


Fig. 10 Visual comparison of salient object detection results. Top three rows, middle three rows, and bottom three rows are from ASD [3, 24], ECSSD [40, 48], and ImgSal [22, 23], respectively. **a** Input images, and **b** ground truth masks. Saliency maps produced by using **c** the proposed CNS model, **d** RPC [25], **e** BMS [52], **f** FES [42], **g** GC [8], **h** HFT [23], **i** PCA [28], **j** RC [7], and **k** TLLT [13]

#	Model	ASD [24,2]			ECSSD [48,40]			ImgSal [22,23]			Average		
		AvgF	MaxF	AdpF	AvgF	MaxF	AdpF	AvgF	MaxF	AdpF	AvgF	MaxF	AdpF
1	AC [1]	.2139	.5107	.5174	.1688	.3766	.3575	.2298	.3807	.3611	.2042	.4227	.4120
2	BMS [52]	.7285	.8555	.8515	.5214	.6302	.6370	.4605	.5401	.4646	.5701	.6753	.6510
3	CA [12]	.4043	.5615	.5569	.3403	.4661	.4314	.3913	.5910	.4801	.3786	.5395	.4895
4	COV [11]	.3413	.6305	.6264	.3347	.5973	.5931	.3485	.4960	.4419	.3415	.5746	.5538
5	FES [42]	.4484	.6859	.6840	.3762	.5951	.5976	.3371	.4557	.4268	.3872	.5789	.5695
6	FT [2]	.4342	.6681	.6677	.2419	.3915	.3775	.2234	.3451	.3380	.2998	.4682	.4611
7	GC [8]	.7474	.8193	.8169	.5118	.5814	.5652	.3381	.3642	.3531	.5324	.5883	.5784
8	GU [8]	.7454	.8164	.8141	.5103	.5774	.5558	.3339	.3646	.3419	.5299	.5862	.5706
9	HC [9]	.6113	.7255	.7009	.3642	.4224	.3894	.2849	.3561	.3238	.4202	.5013	.4714
10	HFT [23]	.4526	.6839	.6806	.3739	.5849	.5652	.4254	.6079	.5129	.4173	.6255	.5862
11	MSS [3]	.4116	.7321	.7369	.2543	.4873	.4864	.2656	.4415	.3807	.3105	.5536	.5347
12	PCA [28]	.5884	.8101	.7953	.4252	.5987	.5778	.4415	.5718	.4679	.4850	.6602	.6137
13	RC [9]	.5192	.7570	.6809	.5766	.6860	.6801	.4048	.4871	.4365	.5002	.6434	.5992
14	RPC [25]	.5762	.8002	.7880	.3757	.5499	.5479	.3400	.4598	.3907	.4306	.6033	.5755
15	SEG [35]	.4305	.6485	.5288	.3840	.4990	.3883	.3096	.4569	.4470	.3747	.5348	.4547
16	SeR [39]	.3975	.5037	.4300	.3179	.3818	.2452	.2855	.4513	.3216	.3336	.4456	.3323
17	SIM [31]	.3162	.4384	.2002	.3080	.3998	.1342	.2497	.4626	.3698	.2913	.4336	.2347
18	SR [17]	.1435	.3964	.3964	.1275	.3469	.3246	.3006	.4324	.3687	.1905	.3919	.3632
19	SUN [53]	.2916	.4402	.3803	.2442	.3522	.2365	.1764	.3198	.2937	.2374	.3708	.3035
20	SWD [10]	.4399	.6434	.6033	.4074	.5700	.4971	.3016	.4787	.4605	.3830	.5640	.5203
21	TLLT [13]	.8270	.8699	.8799	.5832	.6543	.6671	.4512	.4878	.4874	.6205	.6707	.6781
22	CNS	.8204	.8361	.8398	.6191	.6645	.6593	.5902	.6127	.5702	.6765	.7044	.6898
Average		.4950	.6742	.6444	.3803	.5188	.4779	.3404	.4620	.4109	.4052	.5517	.5111

Fig. 11 Statistics of average F_β (AvgF), maximum F_β (MaxF), and F_β using adaptive threshold (AdpF) on three evaluation data sets. The top three scores under each metric are highlighted in red, green, and blue, respectively. See the text for details

detecting salient objects. For example, the BMS model is initially designed for the task of eye fixation prediction. We can see that on average it ranks the third and performs better than most of the salient object detection models evaluated in our experiments.

4.5 Discussions

Although the proposed model performs well on the evaluation data sets, it does fail in some cases. These failures are mainly caused by three visual attributes implicitly used for identifying salient objects: location, color, and size. Figure 12 shows several hard image cases collected from the evaluation data sets. The third row are the color name images annotated by using the RGB colors given in Table 1.

- *Location:* The key idea of BMS is the Gestalt principle based surroundedness, thus the salient regions connected to the image borders would be masked out in the generation of attention maps, as shown in Fig. 12b.
- *Color:* The proposed model originates from BMS, and exploits eleven color name channels for figure-ground segregation. Sometimes, the foreground objects do not directly touch the image borders, but may have very similar colors to the backgrounds. For example, in the third rows of Fig. 12c and d, the RGB colors of the manually labeled salient objects (the horse and the statue) and some background regions (e.g., the valley and the plinth) are almost the same. While salient objects and image borders are connected by background elements, the salient objects are always removed in the generation of attention maps. Moreover, the color statistics based global contrast is introduced in the proposed model. The color similarities between foreground regions and background elements impact the ability of literally popping out salient objects (cf. Fig. 12c and d).

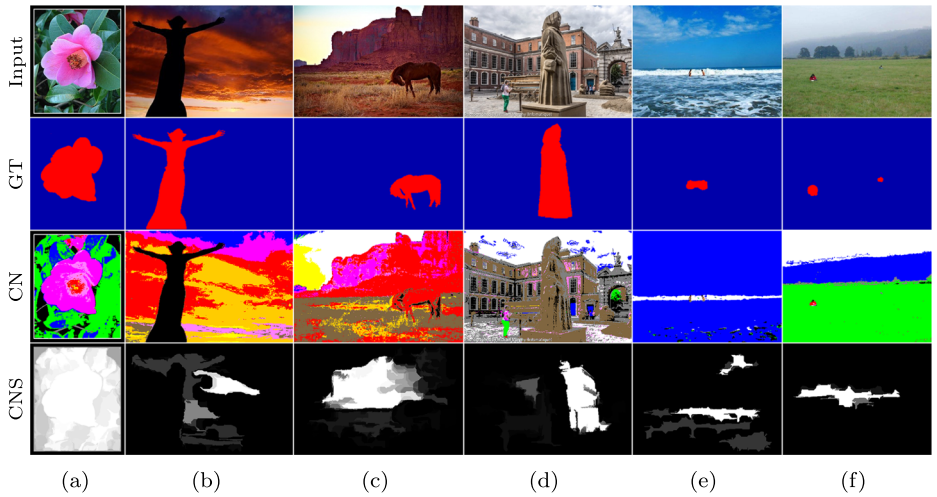


Fig. 12 Hard image cases. Left two columns, middle two columns, and right two columns are from ASD [3, 24], ECSSD [40, 48], and ImgSal [22, 23], respectively. Input: input images. GT: ground truth masks. CN: color name images. CNS: salient object detection results of the proposed model

– *Size:* In the proposed model, some morphological operations, including closing and reconstruction, are used to compute saliency maps. The influences of the parameters ω_c and ω_r have been presented in Fig. 8. These parameters have a substantial impact on the outputs of our model, especially on the ImgSal data set. As Fig. 12e and f show, the manually labeled regions are eroded because the morphological structures are larger than the sizes of these regions.
– Another hard case is caused by the thin artificial borders around some test images, as illustrated in Fig. 12a. When doing the clear-border operation on boolean maps, the proposed model will regard the inner area as a whole region which is surrounded by an enclosed boundary, and does not set any of the foreground pixels to 0. Such a processing mechanism leaves unchanged background elements inside the artificial borders, and results in the failure of figure-ground segregation.

Clearly, the proposed model focuses on the bottom-up image processing technique, and only exploits some low-level image features. Therefore, it fails to highlight the regions that have similar colors to their surroundings. One way to tackle this issue is to invoke more complex visual features. Second, under the definition of surroundedness, the regions connected to the image borders are not enclosed by any complete outer contour. This results in the absence of object-level information in the attention map computation. The above problem can be solved by invoking some background priors and top-down cues. Finally, the proposed model works well for detecting large salient objects, but is not suitable for small ones. It would be interesting to adopt a multi-scale strategy or automatically seek the optimal scale for the detection of different sizes of salient objects.

5 Conclusions

Throughout this paper, we present a salient object detection model based on color name space. Considering the outstanding contribution of color contrast for saliency detection,

a unified framework is constructed to overcome the limitation of the boolean map based saliency. By exploiting several visual features with respect to linguistic color names, we suggest that the model of fusing color attributes provides superior performance over that only based on the surroundedness cue. Moreover, we propose an improved post-processing procedure to uniformly smooth and highlight salient regions, so that the detected salient objects have high and constant intensity levels for the convenience of object segmentation. Experimental results indicate the performance improvement of the proposed model on three evaluation data sets.

With regard to future work, first, we intend to invoke a background measure to handle the salient objects that heavily connected to the image borders. Second, it would be interesting to incorporate more visual features and top-down cues to solve the problem of color confusion between foreground regions and backgrounds. Third, for the morphological structures used in the proposed model, only a fixed value is chosen as the optimal kernel radius, which results in the loss of small salient objects. We have noted that an adaptive radius can effectively address this issue. How to automatically determine the radius size is left to future investigation. Finally, the current version of our MATLAB code is implemented for the purpose of academic research. We further plan to optimize the code to improve the speed performance of the proposed model.

Acknowledgements J. Lou is supported by the Natural Science Foundation of the Higher Education Institutions of Jiangsu Province (No. 19KJB520022), the Science and Technology Special Project of CZIMT (No. 2019-ZXKJ-02), the Cultivation Object of Major Scientific Research Project of CZIMT (No. 2019ZDXM06), the Jiangsu Province Industry University Research Cooperation Project (No. FZ20190200), the Changzhou Key Laboratory of Industrial Internet and Data Intelligence (No. CM20183002), and the QingLan Project of Jiangsu Province (2018). The work of L. Chen, F. Xu, W. Zhu, and M. Ren is supported by the National Natural Science Foundation of China (Nos. 61231014 and 61727802). H. Wang is supported by the National Defense Pre-research Foundation of China (No. 9140A01060115BQ02002) and the National Natural Science Foundation of China (No. 61703209). Q. Xia is supported by the National Natural Science Foundation of China (No. 61403202) and the China Postdoctoral Science Foundation (No. 2014M561654). The authors thank Andong Wang and Haiyang Zhang for helpful discussions regarding this manuscript.

References

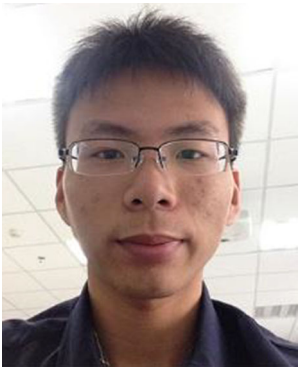
1. Achanta R, Süsstrunk S (2010) Saliency detection using maximum symmetric surround. In: Proceedings of the IEEE international conference on image processing, pp 2653–2656
2. Achanta R, Estrada F, Wils P, Süsstrunk S (2008) Salient region detection and segmentation. In: Proceedings of the international conference on computer vision systems, pp 66–75
3. Achanta R, Hemami S, Estrada F, Süsstrunk S (2009) Frequency-tuned salient region detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1597–1604
4. Berlin B, Kay P (1969) Basic color terms: their universality and evolution. University of California Press, Berkeley
5. Borji A, Frintrop S, Sihite DN, Itti L (2012) Adaptive object tracking by learning background context. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 23–30
6. Borji A, Cheng MM, Jiang H, Li J (2015) Salient object detection: a benchmark. IEEE Trans Image Process 24(12):5706–5722
7. Cheng MM, Zhang G, Mitra NJ, Huang X, Hu SM (2011) Global contrast based salient region detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 409–416
8. Cheng MM, Warrell J, Lin W, Zheng S, Vineet V, Crook N (2013) Efficient salient region detection with soft image abstraction. In: Proceedings of the IEEE international conference on computer vision, pp 1529–1536
9. Cheng MM, Mitra NJ, Huang X, Torr PHS, Hu SM (2015) Global contrast based salient region detection. IEEE Trans Pattern Anal Mach Intell 37(3):569–582

10. Duan L, Wu C, Miao J, Qing L, Fu Y (2011) Visual saliency detection by spatially weighted dissimilarity. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 473–480
11. Erdem E, Erdem A (2013) Visual saliency estimation by nonlinearly integrating features using region covariances. J Vis 13(4):11:1–20
12. Goferman S, Zelnik-Manor L, Tal A (2010) Context-aware saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2376–2383
13. Gong C, Tao D, Liu W, Maybank S, Fang M, Fu K, Yang J (2015) Saliency propagation from simple to difficult. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2531–2539
14. Gonzalez RC, Woods RE, Eddins SL (2004) Morphological image processing. In: Digital image processing using MATLAB. 1st edn. Pearson Prentice Hall, pp 362–365. chap. 9
15. Guo C, Zhang L (2010) A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. IEEE Trans Image Process 19(1):185–198
16. Hofmann T (1999) Probabilistic latent semantic indexing. In: Proceedings of the ACM SIGIR conference on research and development in information retrieval, pp 50–57
17. Hou X, Zhang L (2007) Saliency detection: a spectral residual approach. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–8
18. Itti L, Koch C, Niebur E (1998) A model of saliency-based visual attention for rapid scene analysis. IEEE Trans Pattern Anal Mach Intell 20(11):1254–1259
19. Klein DA, Frintrop S (2011) Center-surround divergence of feature statistics for salient object detection. In: Proceedings of the IEEE international conference on computer vision, pp 2214–2219
20. Koch C, Ullman S (1985) Shifts in selective visual attention: towards the underlying neural circuitry. Hum Neurobiol 4:219–227
21. Lee YJ, Ghosh J, Grauman K (2012) Discovering important people and objects for egocentric video summarization. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1346–1353
22. Li J, Levine MD, An X, He H (2011) Saliency detection based on frequency and spatial domain analysis. In: Proceedings of the British machine vision conference, pp 86:1–11
23. Li J, Levine MD, An X, Xu X, He H (2013) Visual saliency based on scale-space analysis in the frequency domain. IEEE Trans Pattern Anal Mach Intell 35(4):996–1010
24. Liu T, Sun J, Zheng NN, Tang X, Shum HY (2007) Learning to detect a salient object. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–8
25. Lou J, Ren M, Wang H (2014) Regional principal color based saliency detection. PLoS ONE 9(11):e112475:1–13
26. Lou J, Zhu W, Wang H, Ren M (2017) Small target detection combining regional stability and saliency in a color image. Multimed Tools Appl 76(13):14781–14798. https://doi.org/10.1007/s11042-016-4025-7
27. Ma YF, Zhang HJ (2003) Contrast-based image attention analysis by using fuzzy growing. In: Proceedings of the ACM international conference on multimedia, pp 374–381
28. Margolin R, Tal A, Zelnik-Manor L (2013) What makes a patch distinct? In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1139–1146
29. Martin DR, Fowlkes CC, Malik J (2004) Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Trans Pattern Anal Mach Intell 26(5):530–549
30. Mazza V, Turatto M, Umiltà C (2005) Foreground-background segmentation and attention: a change blindness study. Psychol Res 69(3):201–210
31. Murray N, Vanrell M, Otazu X, Parraga CA (2011) Saliency estimation using a non-parametric low-level vision model. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 433–440
32. Niebur E, Koch C (1998) Computational architectures for attention. In: The attentive brain, chap. 9. MIT Press, Cambridge, pp 163–186
33. Palmer SE (1999) Vision science: from photons to phenomenology. The MIT Press, Cambridge
34. Qin C, Zhang G, Zhou Y, Tao W, Cao Z (2014) Integration of the saliency-based seed extraction and random walks for image segmentation. Neurocomputing 129:378–391
35. Rahtu E, Kannala J, Salo M, Heikkilä J (2010) Segmenting salient objects from images and videos. In: Proceedings of the European conference on computer vision, pp 366–379
36. Ren Z, Gao S, Chia Lt, Tsang IWh (2014) Region-based saliency detection and its application in object recognition. IEEE Trans Circ Syst Video Technol 24(5):769–779
37. Rubin E (1958) Figure and ground. In: Readings in perception, pp 194–203
38. Scharfenberger C, Wong A, Clausi DA (2015) Structure-guided statistical textural distinctiveness for salient region detection in natural images. IEEE Trans Image Process 24(1):457–470

39. Seo HJ, Milanfar P (2009) Static and space-time visual saliency detection by self-resemblance. J Vis 9(12):15: 1–27
40. Shi J, Yan Q, Xu L, Jia J (2016) Hierarchical image saliency detection on extended CSSD. IEEE Trans Pattern Anal Mach Intell 38(4):717–729
41. Soille P (1999) Morphological image analysis: principles and applications, 2nd edn. Springer, Berlin
42. Tavakoli HR, Rahtu E, Heikkilä J (2011) Fast and efficient saliency detection using sparse sampling and kernel density estimation. In: Proceedings of the scandinavian conference on image analysis, pp 666–675
43. Tian H, Fang Y, Zhao Y, Lin W, Ni R, Zhu Z (2014) Salient region detection by fusing bottom-up and top-down features extracted from a single image. IEEE Trans Image Process 23(10):4389–4398
44. Treisman AM, Gelade G (1980) A feature-integration theory of attention. Cogn Psychol 12(1):97–136
45. Vincent L (1993) Morphological grayscale reconstruction in image analysis: applications and efficient algorithms. IEEE Trans Image Process 2(2):176–201
46. van de Weijer J, Schmid C, Verbeek J (2007) Learning color names from real-world images. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1–8
47. van de Weijer J, Schmid C, Verbeek J, Larlus D (2009) Learning color names for real-world applications. IEEE Trans Image Process 18(7):1512–1523
48. Yan Q, Xu L, Shi J, Jia J (2013) Hierarchical saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1155–1162
49. Yang C, Zhang L, Lu H (2013) Graph-regularized saliency detection with convex-hull-based center prior. IEEE Signal Process Lett 20(7):637–640
50. Yang C, Zhang L, Lu H, Ruan X, Yang MH (2013) Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3166–3173
51. Zhai Y, Shah M (2006) Visual attention detection in video sequences using spatiotemporal cues. In: Proceedings of the ACM international conference on multimedia, pp 815–824
52. Zhang J, Sclaroff S (2013) Saliency detection: a Boolean map approach. In: Proceedings of the IEEE international conference on computer vision, pp 153–160
53. Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW (2008) SUN: a Bayesian framework for saliency using natural statistics. J Vis 8(7):32: 1–20

Jing Lou received the BE and ME degrees from Nanjing University of Science and Technology, Nanjing, Jiangsu, P.R. China. He is currently a senior engineer with the School of Information Engineering, Changzhou Vocational Institute of Mechatronic Technology. His research interests include image processing, computer vision, and deep learning.

Huan Wang received the PhD degree in pattern recognition and intelligent system from Nanjing University of Science and Technology (NUST), Nanjing, Jiangsu, P.R. China. He is currently an associate professor with the School of Computer Science and Engineering, NUST. His current research interests include pattern recognition, robot vision, image processing, and artificial intelligence.

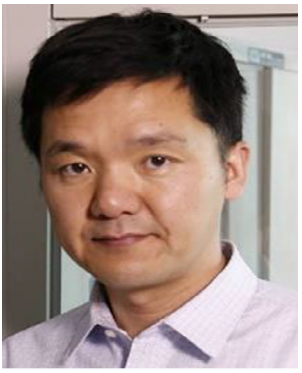


Longtao Chen received the BE degree in computer science and technology from Nanjing University of Science and Technology (NUST), Nanjing, Jiangsu, P.R. China. He is currently working toward the PhD degree in the School of Computer Science and Engineering, NUST. His current research focuses on multi-object tracking.

Fenglei Xu received the BE degree in computer science and technology from Nanjing University of Science and Technology, Nanjing, Jiangsu, P.R. China, in 2014, where he is currently working toward the PhD degree. His research focuses on computer vision.

Qingyuan Xia received the PhD degree in flight vehicle design from Nanjing University of Aeronautics and Astronautics, Nanjing, Jiangsu, P.R. China, in 2013. He is currently a lecturer with the School of Computer Science and Engineering, Nanjing University of Science and Technology, where he worked as a PostDoc from January 2015 to December 2016. His current research interests include environment understanding and navigation technology of intelligent robot, simulation and control of unmanned aerial vehicles.

Wei Zhu received the BE degree in software engineering from Nanjing University of Science and Technology (NUST), Nanjing, Jiangsu, P.R. China. He is currently working toward the PhD degree in the School of Computer Science and Engineering, NUST. His research interests include image processing and deep learning.



Mingwu Ren received the PhD degree in pattern recognition and intelligent system from Nanjing University of Science and Technology (NUST), Nanjing, Jiangsu, P.R. China, in 2001. He is currently a professor with the School of Computer Science and Engineering, NUST. His current research interests include computer vision, image processing, and pattern recognition.