Unsupervised image saliency detection with Gestalt-laws guided optimization and visual attention based refinement.

YAN, Y., REN, J., SUN, G., ZHAO, H., HAN, J., LI, X., MARSHALL, S. and ZHAN, J.

2018
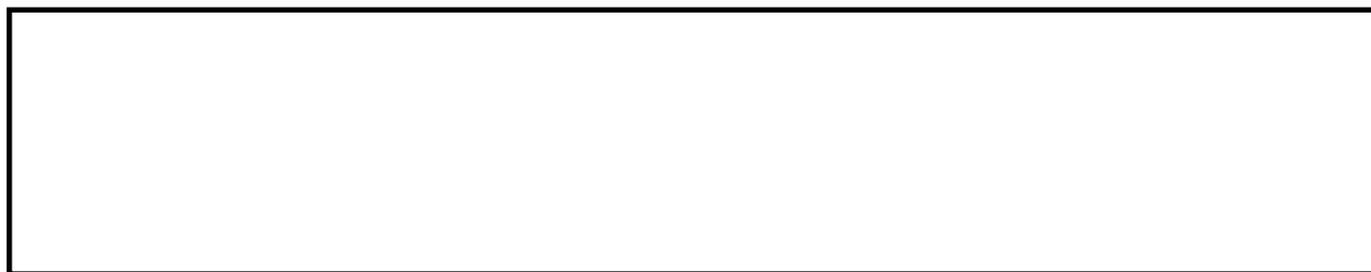
Unsupervised Image Saliency Detection with Gestalt-laws Guided Optimization and Visual Attention Based Refinement

Yijun Yan, Jinchang Ren, Genyun Sun, Huimin Zhao, Junwei Han, Xuelong Li,

Stephen Marshall and Jin Zhan

*J Ren (jinchang.ren@strath.ac.uk), G. Sun (genyunsun@163.com) and H. Zhao (zhaohuimin@gpnu.edu.cn) are joint corresponding authors.

Abstract

Visual attention is a kind of fundamental cognitive capability that allows human beings to focus on the region of interests (ROIs) under complex natural environments. What kind of ROIs that we pay attention to mainly depends on two distinct types of attentional mechanisms. The bottom-up mechanism can guide our detection of the salient objects and regions by externally driven factors, i.e. color and location, whilst the top-down mechanism controls our biasing attention based on prior knowledge and cognitive strategies being provided by visual cortex. However, how to practically use and fuse both attentional mechanisms for salient object detection has not been sufficiently explored. To the end, we propose in this paper an integrated framework consisting of bottom-up and top-down attention mechanisms that enable attention to be computed at the level of salient objects and/or regions. Within our framework, the model of a bottom-up mechanism is guided by the gestalt-laws of perception. We interpreted gestalt-laws of homogeneity, similarity, proximity and figure and ground in link with color, spatial contrast at the level of regions and objects to produce feature contrast map. The model of top-down mechanism aims to use a formal computational model to describe the background connectivity of the attention and produce the priority map. Integrating both mechanisms and applying to salient object detection, our results have demonstrated that the proposed method consistently outperforms a number of existing unsupervised approaches on five challenging and complicated datasets in terms of higher precision and recall rates, AP (average precision) and AUC (area under curve) values.

Y. Yan, J. Ren and S. Marshall are with Department of Electronic and Electrical Engineering, University of Strathclyde, Glasgow, UK (e-mail for corresponding author: jinchang.ren@strath.ac.uk).

G. Sun is with the School of Geosciences, China University of Petroleum (East China), Qingdao, China

H. Zhao and J. Zhan are with the School of Computer Sciences, Guangdong Polytechnic Normal University, Guangzhou, China.

J. Han is with the School of Automation, Northwestern Polytechnical University, Xi'an, China.

X. Li is with Xi'an Institute of Optics and Precision Mechanics, China Academy of Science, Xi'an, China.

1   INTRODUCTION

For human beings, our visual attention system is mainly made up by both bottom-up and top-down attention mechanisms that enable us to allocate to the most salient stimuli, location, or feature that evokes the stronger neural activation than others in the natural scenes [5-7]. Bottom-up attention helps us gather information from separated feature maps e.g. color or spatial measurements, which is then incorporated to a global contrast map representing the most salient objects/regions that pop out from their surroundings [11]. Top-down attention modulates the bottom-up attentional signals and helps us voluntarily focus on specific targets/objects i.e. face and cars [15]. However, due to the high level of subjectivity and lack of formal mathematical representation, it is still very challenging for computers to imitate the characteristics of our visual attention mechanisms. In [11], it is found that the two attentional functions have distinct neural mechanisms but constantly influence each other to attentions. To this end, we aim to build a cognitive framework where separated model for each attentional mechanism is integrated together to determine the visual attention refer to the salient object detection.

To extract features at the bottom level, color plays an important role since it is a central component of the human visual system, which also facilitates our capability for scene segmentation and visual memory [22]. Color is particularly useful for object identification as it is invariant under different viewpoints. We can move or even rotate an object, yet the color we see seems unchanged due to the light reflected from the object into the retina remains the same. As a result, the salient regions/objects can be easily recognized intuitively for their high contrast to the

| Image | LMLC[1] | HDCT[8] | RC[3] | GMR[12] | Proposed | Ground truth |
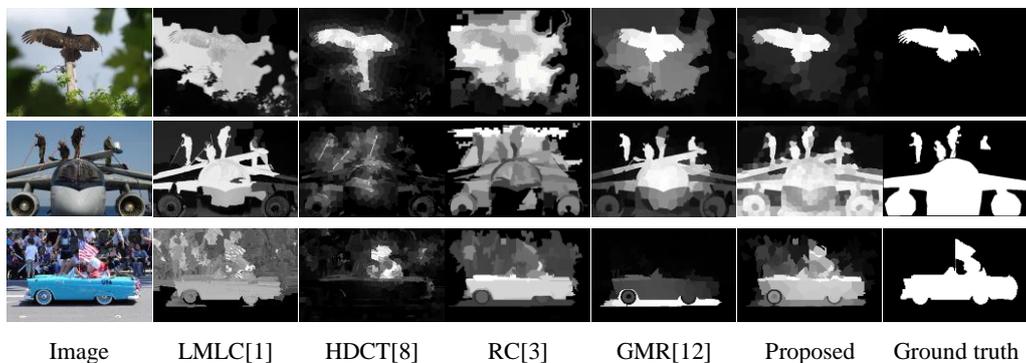
Fig. 1. Three examples of salient objects.

surrounding background.

In addition to color features, our visual perception system is also sensitive to spatial signals, as the retinal ganglion cells can transmit the spatial information within natural images to the brain [25]. As a result, our human beings pay more attention to the objects and regions not only with dominant colors but also with close and compact spatial distributions. Therefore, the main objective of saliency detection is to computationally group the perceptual objects on the base of the way how our human visual perception system works.

Although color and spatial features have been widely used for salient object detection, the efficacy can still be fragile, especially in dealing with large objects and/or complicated background in the scenes[23]. The salient object often cannot be extracted as a whole (see examples in Fig. 1), though it is still relatively easily for our HVS to identify the full range of the salient objects. This shows a gap between existing approaches to an ideal one that can better exploit the potential of our HVS for more accurate salient object detection. To this end, we propose a Gestalt-law guided cognitive approach to calculate bottom-up attention. As gestalt-laws can characterize the capabilities of HVS to yield whole forms of objects from a group of simple and even unrelated visual elements [27], e.g. edges and regions, we aim to employ these laws to guide/improve the process of salient object detection.

For modelling top-down attention, Al-Aidroos et al [28] proposed a theory named 'background connectivity' to describe the stimulus-evoked response of our visual cortex. It is found that focus on the scenes rather than objects may increase the background connectivity. Inspired by this theory, we employed a robust background detection model to represent the background connectivity of top-down attention in the images as post-processing to further refine the saliency maps detected using gestalt-laws guided processing.

Fig. 1 shows several examples in which the salient objects contain poor color and/or spatial contrasts. As such, conventional approaches either fails to detect the object as a whole or results in massive false alarms. Within the proposed cognitive framework, salient objects can be successfully detected whilst the false alarms are significantly suppressed. Descriptions of the proposed salient model and its implementation are detailed in Sections 3-4.

The main contributions of this paper can be highlighted as follows:

1) We propose gestalt laws guided optimization and visual attention based refinement framework (GLGOV) for unsupervised salient object detection, where bottom-up and top-down mechanisms are combined to fully characterize HVS for effective forming of objects in a whole;

2) We introduce a new background suppression model guided by the Gestalt law of figure and ground, where superpixel-level color quantization and adaptive thresholding are applied to determine object-level foreground and background for the calculation of the background correlation term and the spatial compactness term to further suppress the background and highlight the saliency objects;

3) We have carried out comprehensive experiments on five challenging and complex datasets and benchmarked with eight state-of-the-art saliency detection models, where useful discussions and conclusions are achieved.

The rest of this paper is organized as follows. Section 2 summarizes the related work on saliency detection. The proposed framework by combining bottom-up and top-down HVS mechanisms for saliency detection is presented in Section 3, where the implementation detail is discussed in Section 4. Section 5 presents the experimental results and performance analysis. Finally, some concluding remarks are drawn in Section 6.

2   RELATED WORK

In the past decades, a number of salient object detection methods have been developed to identify salient regions in terms of the saliency map and capture as much as possible human perceptual attention. In general saliency detection methods can be categorized into two classes, i.e. supervised and unsupervised approaches. Most supervised methods including those using deep learning [29-33] are able to obtain good saliency maps, where high performance computers even with particular graphic process units (GPU) are needed to cope with the lengthy training time. In addition, supervised methods may also suffer from lack of generality, especially when the training samples are limited and/or insufficiently representative. With deep learning, this drawback seems can be somehow overcome [32], yet at a cost of a large amount of data requested for training to learn the prior knowledge. On the contrary, it seems our human vision system can guide us to easily detect and recognize objects under complex scenes without supervision [34]. To this end, in this paper we focus mainly on unsupervised saliency detection.

In a recent benchmark survey [35], quite a few unsupervised saliency detection methods are summarized and assessed, where the two main objectives of saliency detection are fixation prediction [4, 10] and salient object detection [36-39]. In fixation prediction, it aims to predict eye's gaze or motion through detecting sparse blob-like salient regions [40], whilst salient object detection is to detect the salient objects/regions in the scene [41]. According to the survey [35], much more salient object detection methods are proposed than those using eye fixation prediction,

possibly due to their contributions to a wide range of applications including content-based image retrieval [42-44], image/video compression [45-47], image quality assessment [48-50], region of interest segmentation [51-53], and object detection [54-56], etc.

Inspired by a biologically plausible architecture [6] and the feature integration theory [57], Itti et al [4] proposed an epic saliency detection model in 1998. With multiple image features extracted including luminance, color and edge orientation, the saliency map is generated by using center-surround difference across these features. In the following two decades, quite a few landmark saliency models are developed, which are briefly reviewed below.

Depending on whether a salient object is detected from pixels or regions, saliency detection techniques can be further categorized into two groups, i.e. pixel based and region based. Herein the main difference between the two groups is whether the image is segmented into regions for saliency detection, using either color quantization or pixel clustering. In [9], a contrast based saliency map is proposed, where the color difference between the pixel and its neighbors is determined to extract the attended areas using the fuzzy theory [58]. In [10] a biologically plausible bottom-up visual saliency model is presented based on the Markovian approach and mass concentration algorithm. More recently, an efficient method is introduced in Achanta et al. [16] to build high quality saliency maps using low-level features such as luminance and color in the $L*a*b*$ color space. In [18], a salient region is detected by using the color difference between the pixels and their average value in the image, again in the $L*a*b*$ color space. In [19], instead of treating the whole image as the common surround for any given pixel, the saliency map is defined by using color difference between the given pixel and a local symmetric surround region. In Cheng et al [24], a global contrast based method is proposed to determine the saliency value. The color is quantized into a number of bins in the $L*a*b*$ color space with the global color contrast measured between color bins. Furthermore, a color space smoothing process is also introduced to reduce quantization artefacts before assigning similar saliency value to similar color bins.

Although the aforementioned approaches are found to produce relatively good results on saliency detection, their robustness is limited when extending to large datasets due to increasing complexity of the scenes, especially the variations in terms of spatial size and layout between the salient objects and the image background. The reason here is, except for the color contrast, spatial contrast is also an important perception in our human visual system, regardless the extremely high computational cost for pixel-level saliency computation. To this end, region-based contrast and saliency detection has become increasingly popular in recent years, especially using the superpixel based approach. In

[24], a region-level saliency map is proposed based on both color and spatial difference across the regions, where the spatial prior is used to highlight the salient regions. In [59], a contrast-based saliency estimation is proposed, where a given image is segmented into a number of homogeneous regions by using superpixel. The contrast and spatial distribution of these regions are measured and smoothed by using high-dimensional Gaussian filters for saliency detection. In [14], a superpixel based saliency detection method is proposed, where color and spatial contrast across the superpixels are used for efficient saliency detection. In Kim et al [8], high-dimensional color transform is applied to over-segmented images for saliency detection.

In general, the whole process of bottom-up saliency detection can be divided into at least three stages, i.e. pre-processing, feature based salient map generation, and post-processing. Apparently feature based salient map generation is the key in saliency models, where various color and spatial features are extracted and measured in determining the saliency maps. The pre-processing is often for spatial and illumination normalization, image enhancement and image segmentation (only for region-based approaches). The post-processing, on the contrary, serves mainly for normalization and/or fusion of saliency maps, where object prior is widely used in region-based approaches. One optional stage is to extract the binary template of the salient object via thresholding the salient map, where histogram based adaptive thresholding such as Otsu's approach [60] is commonly used.

In Table I, some typical unsupervised saliency detection approaches are summarized for comparison in terms of the features used and any adopted pre-processing and post-processing stages. First, color and spatial information is the

Table I
Overview of some popular unsupervised saliency detection models

	Method	Pre-processing	Features for initial saliency map generation	Post-processing to refine the saliency map
Pixel-level saliency detection	IT[4]	No	Color, Intensity, Orientation	Normalization of saliency map
	MZ[9]	Image resizing, Color quantization	Color	Refinement via Fuzzy growing
	GB[10]	No	Color, Orientation, Intensity	Normalization of saliency map
	SR[13]	Log spectrum	Intensity	No
	AC[16]	No	Color, Luminance	Fusion of multi-saliency maps
	FT[18]	Gaussian filter	Color, Luminance	No
	MSS[19]	No	Color	No
	SEG[21]	No	Color	No
	HC[24]	Color quantization	Color, Luminance	No
Region-level saliency detection	RC[3]	Color quantization, Graph-based image segmentation	Color, Luminance, Spatial	Object prior with color refinement and hard constraints
	SP[14]	Color quantization, Superpixel clustering	Color, Spatial	Color and spatial refinement
	LMLC[1]	Superpixel clustering	Color, Spatial, Intensity	Color refinement
	HDCT[8]	Superpixel clustering	Location, Color, Texture, Shape	Object prior via Spatial refinement

most widely used features due to their importance in visual psychology [22, 25]. However, spatial features are excluded in some pixel-based approaches. For region-based approaches, color quantization and graph or superpixel based clustering is normally used. Combination of color and spatial features are then employed for saliency map determination and refinement.

As shown in Table I, the principle of gestalt laws has been reflected in many existing approaches. This includes not only the color and spatial features in representing the HVS but also their applications in other key stages. Examples herein can be found in homogeneity based color quantization and pixel clustering in the pre-processing, similarity based feature measurement in saliency map generation and object-prior based grouping in post-processing. In fact, the concept of gestalt laws was first introduced as a series of mechanisms to explain the human visual perception, dated back to 1920s in [61, 62], where Gestalt law based detection is proved to fit the human perception [63]. Starting from low-level features of the salient stimuli that influence perception at the bottom level way up to high level cognition, gestalt can be naturally used to define bottom-up objects [64, 65]. Although the gestalt theories are gradually used in saliency detection models explicitly [66-68] or implicitly [59, 69], there is no specified structure to regulate the relationship between Gestalt law and saliency models. How these laws can be systemically applied for salient detection remains unexplored. As a result, we aim to further explore various aspects of gestalt laws when applying in a bottom-up model and also combined with the top-down model for feature contrast map generation, which is detailed in section 4.

3   THE PROPOSED GLGOV FRAMEWORK FOR UNSUPERVISED SALIENCY DETECTION

A new saliency detection framework inspired by the Gestalt laws of HVS is proposed. The proposed framework contains six main modules, i.e. homogeneity, similarity and proximity, figure and ground, background connectivity, two stage refinement and performance evaluation. The overall diagram of our saliency detection framework is illustrated in Fig. 2 where corresponding gestalt laws and visual psychology used in different modules are specified and also detailed below.

The homogeneity module aims to group or cluster pixels into regions on the basis of human visual perception [62, 70]. According to Gestalt law of homogeneity and the gestalt perceptual organization theory [62], if pixels have similar intensity, color, orientation or other features, they should be treated as homogeneity regions. To this end, there
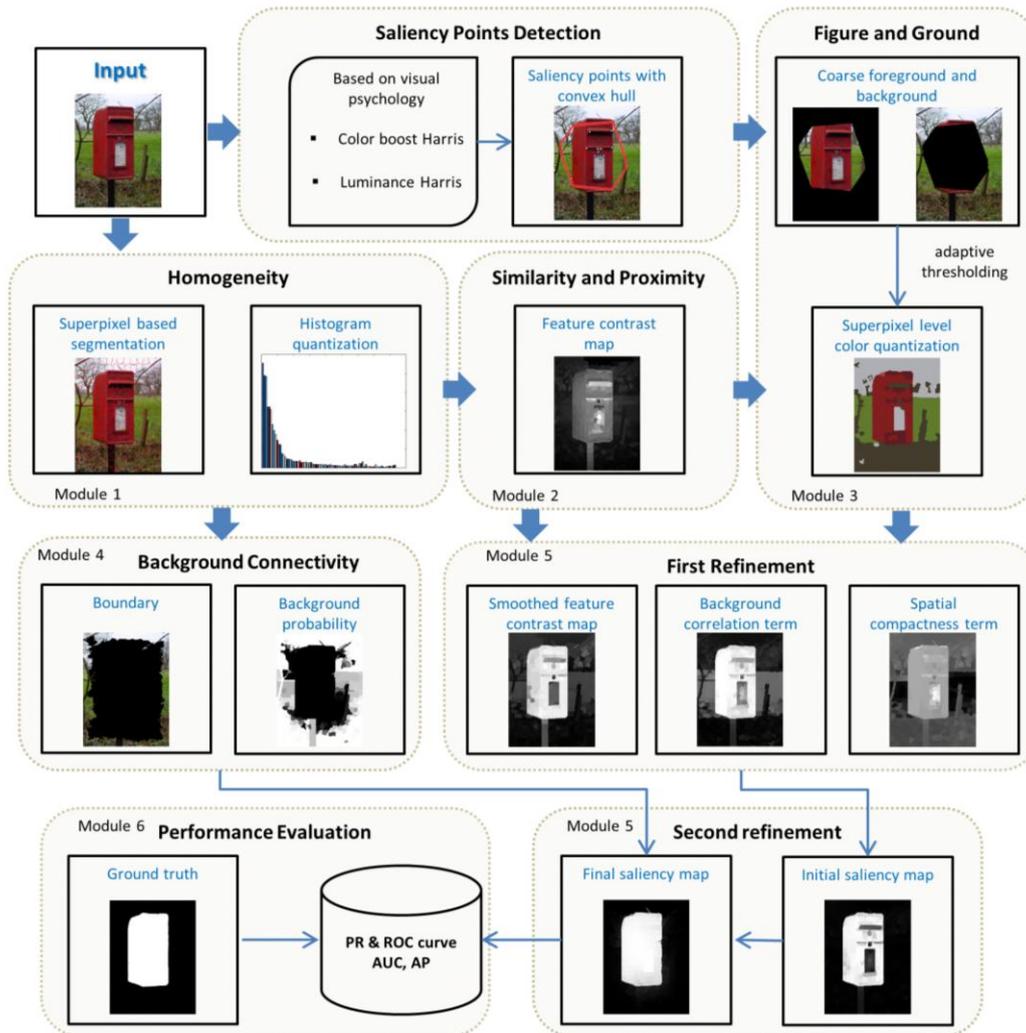
Fig. 2. The proposed framework in six modules guided by various gestalt laws (specified in brackets) and visual psychology.

can be a number of homogeneity regions extracted in one image. Specifically, the simple linear iterative cluster (SLIC) method [71] is used to separate the image into regions namely superpixels, where the color quantization is applied to reduce the number of colors based on color homogeneity.

In similarity and proximity module, color contrast and spatial contrast across superpixels are measured for extraction of the feature contrast map and smoothing process. Based on Gestalt laws of similarity and proximity, elements/objects tend to be perceived as a whole in cognition if they are close enough to each other and/or share similar appearance of visual features. In other words, if a superpixel with a low saliency value is surrounded by those with high saliency values, this superpixel should be assigned with a high saliency value as determined by its

neighboring superpixels provided that their colors are similar to each other. As such, a smoothing process should be introduced to refine the extracted saliency map, which will be defined in two stage refinement.

The saliency point detection module is to coarsely split the whole image into foreground and background based on a convex hull formed by the detected saliency points. Inspired by human visual psychology [22, 25], the greyscale image and color image are treated differently. Herein, we choose luminance-based operator for greyscale images and Color Boosted Harris (CBH) operator for color images for detection of saliency points [72]. As this module is relative quite straightforward, it is not detailed in this paper.

In figure and ground module, Gestalt law of figure and ground is applied to suppress the background and highlight the foreground objects. According to HVS, human's perception is comprised by objects and their surrounding background under observation, where salient objects can be automatically highlighted from the suppressed background. To this end, superpixel level color quantization, adaptive thresholding and a saliency points detection method [72] are well combined together for background suppression.

The background connectivity module is to calculate the background probability of each super-pixels. Inspired by background connectivity theory [73], attention to the scene or background of the image will increase the background connectivity in our visual cortex. Herein, we employ a component from the background detection method [20] to represent the background connectivity mathematically.

In the two-stage refinement module, the first refinement helps to smooth Gestalt law guided saliency maps by considering all superpixels from the foreground or background as well as fusion with the feature contrast map and background suppression map. The result from the first refinement will be fused with the background probability map derived from the background connectivity module and further smoothed in the second refinement process. Finally, the final saliency map is determined.

In the performance evaluation module, comprehensive experiments are carried out against a number of state-of-the-art methods on several widely used publicly available databases. Some widely used evaluation criteria such as PR curve, ROC curve, AUC and AP value are used for quantitative assessment. Detailed results are reported in Section 5.

4 IMPLEMENTATION DETAIL OF THE PROPOSED GLGOV FRAMEWORK

In this section, the implementation of the proposed saliency detection framework is detailed in five stages, i.e. homogeneity, similarity and proximity, figure and ground, background connectivity and two-stage refinement below.

4.1 Homogeneity

For a color image in three channels, it usually contains thousands of pixels with up to 256^3 possible values, where pixel-based saliency detection suffers from high computational cost. Therefore, we use two operations to reduce the dimension of the image and simplify the problem. First, based on Gestalt law of homogeneity, similar pixels spatially close to each other in the image can be considered as a whole. To this end, we use SLIC [71] to segment the image into a number of regions called superpixels which generally has inner color consistency and a compact shape with decent boundary adherence. As suggested by [8], we set the number of superpixels n to 500. Secondly, using the same scheme as suggested in [14], we uniformly quantize each color channel into q bins in order to reduce the number of colors to generate a new image histogram H with q^3 bins. Herein we define $q=12$ instead of 16, as this change has little effect to the experimental results yet the number of colors and associated computation cost has been significantly reduced, i.e. proportionally to the reduction from 16^3 to 12^3.

From the quantized image, bins of dominant colors are selected to cover more than 95% of the image pixels. The rest color bins that cover less than 5% pixels are merged into the selected dominant color bins based on minimum Euclidean distance criterion. An example of the color quantization is shown in Module 4, Fig. 2. After superpixel-based image segmentation and color quantization, each superpixel $S_i (i = 1,2, \dots n)$ has its histogram H_i normalized by $\sum_{m=1}^{k} H_i(m) = 1$, where k is the index of the color bins.

4.2 Similarity and Proximity

In human vision perception, usually we are aware of the image regions that have high contrast to their surroundings. According to the Gestalt law of proximity and similarity, HVS tends to perceive elements that are close or similar to each other in a group. This actually explains that, at the superpixel-level, those image regions are normally grouped by spatial closeness and color similarity of superpixels. In addition, those superpixels should have similar level of low inner contrast and highly different from the surrounding superpixels. In this section, we presented in detail the proposed saliency object detection method based on color and spatial contrast at superpixel level.

First, for two superpixels S_i and S_j, their color distance $C_d(S_i, S_j)$ is defined as the Euclidean distance between the mean colors of the two superpixels and scaled within the range of [0,1]. Here the color distance C_d measures the similarity of the color appearance of the two superpixels.

In addition to the color based similarity, the gestalt law of proximity for the two superpixels S_i and S_j is denoted as a spatial proximity distance $P_d(S_i, S_j)$, which is defined as the Euclidean distance between the centroids of the two superpixels. The larger the $P_d(S_i, S_j)$ is, the higher proximity between S_i and S_j is. The spatial proximity distance can be further normalized within [0, 1] as follows:

$$P_{d_{norm}}(S_i, S_j) = \frac{P_d(S_i, S_j) - P_{d\_max}}{P_{d\_min} - P_{d\_max}} \tag{1}$$

where P_{d_min} and P_{d_max} refer respectively to the minimum and the maximum of all possible values of P_d as determined from superpixels of the given image.

Based on the defined color distance C_d and the normalized spatial proximity distance P_{d_norm}, the global color contrast for a superpixel S_i is defined by

$$C_G(S_i) = \sum_{j=1}^{n} A_j \cdot P_{d\_norm}(S_i, S_j) \cdot C_d(S_i, S_j) \tag{2}$$

where n is the number of superpixels, and A_j is the normalized area of the superpixel S_j where a larger one contributes more to the global color contrast value of S_i, where the sum of A_j is 1.

Similarly, we define the global spatial contrast value for a superpixel S_i below:

$$S_G(S_i) = \frac{\sum_{j=1}^{n} P_{d\_norm}(S_i, S_j) \cdot C_I(S_i, S_j) \cdot L_j}{\sum_{j=1}^{n} L_j \sum_{j=1}^{n} P_{d\_norm}(S_i, S_j) \cdot C_I(S_i, S_j)} \tag{3}$$

where L_j is the weight of the spatial layout, which is defined as the minimum distance from the centroid of superpixel S_j to any of the four image borders; C_I is the inter superpixel color similarity defined below, where H_i and H_j refer respectively to the color histogram of the two superpixels S_i and S_j with m as the index for the k color bins.

$$C_I(S_i, S_j) = \sum_{m=1}^{K} ((1 - |H_i(m) - H_j(m)|) * \min(H_i(m), H_j(m))) \tag{4}$$

Unlike conventional histogram intersection method, we introduce the term $(1 - |H_i(m) - H_j(m)|)$ rather than $\min(H_i(m), H_j(m))$ to measure the similarity of two histogram bins based on their frequencies. For example, given

two pairs of histogram bins: $H_i(1) = 0.3, H_j(1) = 0.2$, and $H_i(2) = 0.8, H_j(2) = 0.2$, conventional histogram intersection method will produce the same result of 0.2, which fails to reflect the difference between the histogram bins. On the contrary, in our modified definition, the similarity terms for the two cases become 0.9 and 0.4, respectively, which improves the similarity measurement and makes it more consistent to human perceptions than conventional histogram intersection approach.

For simplicity, we define ICS as a combined measurement of color similarity and spatial proximity between two superpixels S_i and S_j:

$$ICS(S_i, S_j) = P_{d\_norm}(S_i, S_j) \cdot C_I(S_i, S_j) \tag{5}$$

Accordingly, Eq. (3) can be simplified as

$$S_G(S_i) = \frac{\sum_{j=1}^{n} ICS(S_i, S_j) \cdot L_j}{\sum_{j=1}^{n} ICS(S_i, S_j)} \tag{6}$$

Based on both the global color contrast C_G and the global spatial contrast S_G, the feature contrast value for superpixel S_i is defined as

$$FC(S_i) = C_G(S_i) \cdot S_G(S_i) \tag{7}$$

Although a salient object usually has a high contrast to the background, it may contain some low-contrast parts. As a result, the salient object cannot be detected as a whole as these low-contrast parts will be missed due to their small sizes and low saliency values. To tackle these problems, a smoothing process (detailed in 4.5) is applied to filter such superpixels.

4.3   Figure and Ground

Although the image background can be very complicated and even similar to the foreground, in most cases people can still recognize the salient objects without difficulty, and this can be explained by the Gestalt law of figure-ground that human's perception is the fusion of observed objects and their surrounding background. To this end, we can enhance the contrast of salient superpixels for their easier detection by suppressing the saliency value for the superpixels in the background. Based on gestalt law of similarity and proximity, salient objects in superpixel level image are composed of several superpixels with similar color appearance and close spatial distance. This actually indicates that salient objects are formed by several superpixels in a compact manner yet the distribution of background superpixels tends to be more dispersive. Based on this assumption, in this subsection, we introduce a novel

background suppression model, the spatial distribution of superpixels is utilized to highlight the objects and suppress the background.

According to the dominant color theory [74], images are usually quantized into up to 8 dominant colors. Similarly, for a given superpixel level image, the colors of all superpixels can also be divided into 8 coarse clusters in the $L*a*b*$ color space, where each of the three individual color components is divided into two parts by a threshold $T^{(.)}$. For each of the eight color clusters, the extracted dominant color is denoted as $C_i = [\bar{x}_i^L, \bar{x}_i^A, \bar{x}_i^B]$, $i \leq 8$, where $\bar{x}_i^{(.)}$ is the average value in one of the three color components $L*$, $a*$ and $b*$, and C_i will be used as the center for the corresponding color cluster.

Herein the color boosted Harris point (CBHP) operator [72] is employed to determine the coarse regions of foreground and background. For color or greyscale images, the CBHP operator and luminance operator are respectively applied to detect the salient points, where a convex hull is computed to enclose all these salient points. All superpixels within the convex hull are defined as the foreground and the remaining superpixels as the background (see for example in Fig. 2 module 4). Let $MP_{G_f}^{(.)}$ and $MP_{G_b}^{(.)}$ denote the mean color (with three components) of the foreground G_f and the background G_b. To maximize the color contrast, in each color component the threshold $T^{(.)}$ is adaptively defined as $T^{(.)} = \frac{1}{2}\left(MP_{G_f}^{(.)} + MP_{G_b}^{(.)}\right)$.

With the adaptively determined eight dominant colors above, we merge similar colors for simplicity below. We iteratively examine each pair of the extracted dominant colors and merge them if their Euclidean distance is less than a threshold T_d. For each image, the threshold is adaptively decided as half of the Euclidean distance between $MP_{G_f}^{(.)}$ and $MP_{G_b}^{(.)}$, and the upper limit of T_d is set as 15.

Let C_i and C_j be two color clusters, they can be merged by the weighted average agglomerative procedure [75]:

$$C^{(.)} = \rho C_i^{(.)} + (1 - \rho) C_j^{(.)}$$

$$\rho = \frac{p_i}{p_i + p_j}$$

(8)

where p_i and p_j denote respectively the numbers of pixels in C_i and C_j. For each superpixel-level image, we can usually extract 3-8 dominant colors, located in background or foreground. In other words, we may have up to 16 color clusters obtained, where half of them are from the foreground group and others from the background group. As the

spatial information is excluded in color quantization, within each color cluster the superpixels are not necessarily spatially grouped together. This means that each color cluster may contain several separated objects composed of one or more superpixels. As a result, the superpixel-level image has now become object-level image.

For each color cluster C_k in either G_f or G_b, denote $M(C_k)$ as its the geographic center. The spatial compactness term of C_k is defined as the average distance between $M(C_k)$ and each object O_i it contains:

$$Sd_f(C_k) = \frac{1}{N_k} \sum_{\forall O_i \subset C_k \subset G_f} \|M(O_i) - M(C_k)\| \tag{9a}$$

$$Sd_b(C_k) = \frac{1}{N_k} \sum_{\forall O_i \subset C_k \subset G_b} \|M(O_i) - M(C_k)\| \tag{9b}$$

where $k \leq 8$, N_k denotes the number of objects in the cluster C_k and $M(O_i)$ refers to the geographic center of the object O_i.

Similar to $P_{d_{norm}}$ in Eq. (1), Sd can be scaled into [0, 1] by

$$N_{Sd}(C_k) = \frac{Sd(C_k) - Sd_{max}}{Sd_{min} - Sd_{max}} \tag{10}$$

where $Sd(C_k)$ is either $Sd_f(C_k)$ or $Sd_b(C_k)$; Sd_{min} and Sd_{max} refer respectively to the minimum and the maximum of all spatial compactness values of Sd.

By applying the spatial compactness, the salient objects can be highlighted. Herein we introduce a new background correlation term to further suppress the background and enhance the salient objects. This term is used to measure the connection degree between an object O and the image borders, where a large value indicates a high degree of overlapping between O and the image borders. As a result, the object O will be more likely be classified into the background as we assume the salient object should be away from the image borders. The background correlation term is defined as

$$B_c(S_i) = \frac{N_b}{N_c}, \quad S_i \subset O \tag{11}$$

where N_c denotes the total number of superpixels in O, among which N_b is the number of superpixels in O that directly connects to the image borders. Using the maximum and the minimum of all possible values of B_c, denoted as B_{c_max} and B_{c_min}, we can obtain normalized B_c within [0,1] below.

$$\overline{B_c}(S_i) = \frac{B_c(S_i) - B_{c\_max}}{B_{c\_min} - B_{c\_max}} \qquad (12)$$

4.4 Background connectivity

In this subsection, an effective method of background detection [20] is employed to extract the background connectivity in a numerical value by determining the probability of each superpixels being the background. For a given superpixel S_i, the associated background probability can be determined by:

$$BP(S_i) = \frac{Len(S_i)}{\sqrt{A(S_i)}}, \qquad (13)$$

$$A(S_i) = \sum_{j=1}^{n} exp\left(-\frac{d_{geo}^2(S_i, S_j)}{2\sigma^2}\right), \qquad (14)$$

$$Len(S_i) = \sum_{j=1}^{n} exp(-\frac{d_{geo}^2(S_i, S_j)}{2\sigma^2}) \cdot \delta(S_j). \qquad (15)$$

where σ is set to 10, $\delta(\cdot)$ is 1 for superpixels on the image boundary and 0 otherwise as suggested in [65], and n is the number of superpixels. $d_{geo}(S_i, S_j)$ is the geodesic distance between any two superpixels, which is calculated by accumulating the weights d_{app}, measured using the Euclidean distance along the shortest path formed by a sequence of adjacent superpixels between S_i and S_j on the graph, and l is the total number of superpixels in the determined path.

$$d_{geo}(S_i, S_j) = \min_{S_1 = S_i, S_2, \dots, S_l = S_j} \sum_{i=1}^{l-1} d_{app}(S_i, S_{i+1}) \qquad (16)$$

To scale BP within [0,1], the parameter σ_{BP} is set to 1, and the normalization process is defined as:

$$NBP(S_i) = 1 - exp(-\frac{BP^2(S_i)}{2\sigma_{BP}^2}) \qquad (17)$$

4.5 Two-stage refinement

In the first stage refinement, the feature contrast map, spatial compactness term and background correlation terms will be smoothed separately and fused together to generate initial saliency map.

For each superpixel, its saliency value is replaced using the weighted average of the saliency values of its surrounding superpixels. Similar to image filtering, the process applied to the superpixel level image can also smooth

the saliency values for more consistent detection of salient objects. In [24], a linear varying smoothing operator is employed to smooth the color contrast in the image. However, this smoothing procedure fails to address the spatial factor. In our approach, we improve this procedure with Gestalt law of similarity and proximity, and apply it to superpixel level image filtering. For robustness, for a given superpixel S_i, we choose $k = n/8$ nearest neighbors for weighted average to refine the feature contrast value of S_i locally as follows:

$$RFC(S_i) = \frac{1}{(k-1)T} \sum_{j=1}^{k} (T - ICS(S_i, S_j)) \cdot FC(S_i) \cdot A_j \tag{18}$$

where $T = \sum_{j=1}^{k} ICS(S_i, S_j)$ is the sum of color similarity and spatial proximity between S_i and its k nearest neighbors.

For the input image in Fig. 3(a), Fig. 3(b) shows the contrast-based saliency map obtained from Eq. (7). Due to the shadow-caused low contrast parts in the flower, the saliency map actually contains many low saliency valued superpixels. This will inevitably affect the successful detection of the salient object from the image. As a result, the smoothing process and
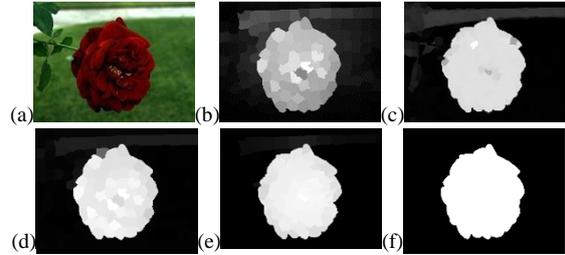
Fig. 3. An example of proposed saliency computation: (a) The original image, (b) Feature contrast map, (c) background suppression map, (d) first-stage refinement, (e) final saliency map after second-stage refinement, and (f) the ground truth map.

background suppression model is applied with the results shown in Fig. 3(c). After smoothing, the saliency map has been improved in several ways. First, the low saliency values from the salient object have been enhanced. Second, the saliency values for all the superpixels become more consistent, which are actually raised. This shows that the proposed smoothing procedure can not only filtering low-contrast defects but also normalize the overall saliency values. Consequently, the quality of the adjusted saliency map is significantly improved.

For all the superpixels in C_k, they will be assigned the same spatial compactness term $N_{sd}(C_k)$ as determined in Section 4.3. The lower this value is, the higher likeness the salient object it has. Although $N_{sd}(C_k)$ can measure the spatial compactness of all superpixels within C_k, it cannot differ between them even for incorrectly detected foreground superpixels. To overcome this drawback, using the similar process in Eq. (18), the Gestalt laws of similarity and proximity are applied to smooth the spatial compactness term for each superpixel S_i as follows:

$$R_{Sd}(S_i) = \frac{1}{(n-1)T} \sum_{j=1}^{n} \Big(T - ICS(S_i, S_j) \Big) \cdot N_{Sd}(S_j), S_j \subset C \tag{19}$$

where again n is the number of superpixels, $T = \sum_{j=1}^{n} ICS(S_i, S_j)$ is the sum of color similarity and spatial proximity between S_i and other superpixels. Note that, in global refinement all superpixels rather than neighboring superpixels are selected for weight average because the spatial compactness term has been extended from superpixel-level to object-level.

In addition, the background correlation term can also be globally refined by

$$R_{B_c}(S_i) = \frac{1}{(n-1)T} \sum_{j=1}^{n} \Big(T - ICS(S_i, S_j) \Big) \cdot \overline{B_c}(S_j), S_j \subset O \tag{20}$$

Finally, the background suppression map is determined by using the conjunction of both the spatial compactness term $R_{Sd}(S_i)$ and background correlation term $R_{B_c}(S_i)$ below:

$$OP(S_i) = R_{Sd}(S_i) \cdot R_{B_c}(S_i) \tag{21}$$

Based on Eq.(18) and Eq. (21), our initial saliency map is formed by

$$ISA(S_i) = RFC(S_i) \cdot OP(S_i) \tag{22}$$

In the second refinement stage, the initial saliency map and the background probability map are fused by adopting a cost function [20] to optimize the whole procedure. Let FSA_i denote the final saliency value determined for the superpixel S_i, the cost function is given by:

$$J(FSA_i) = \min_{FSA_i} \left[\sum_{i=1}^{n} NBP(S_i) \cdot FSA_i^2 + \sum_{i=1}^{n} ISA(S_i) \cdot (FSA_i - 1)^2 + \sum_{i,j=1}^{n} \omega_{i,j} \cdot \big(FSA_i - FSA_j \big)^2 \right], \tag{23}$$

$$\omega_{i,j} = exp(-\frac{d_{app}^2 (S_i, S_j)}{2\sigma^2}) + \mu \tag{24}$$

Note that σ is set to 10 as defined in Eq. (14). As seen in Fig. 3 (d-e), by adding the two-stage refinement, the salient object can be significantly highlighted whilst the saliency value for the background is effectively suppressed.

5   EXPERIMENTAL RESULTS

For performance evaluation of our proposed saliency detection method, in total 10 state-of-the-art algorithms are used for benchmarking, as listed below by the first letter of the name of methods. They are selected for two main

reasons, i.e. high citation and wide acknowledgement in the community and/or newly presented in the last 3-5 years. Introduction to the datasets and criteria used for evaluation as well as relevant results and discussions are presented in detail in this section.

- Bayesian saliency via low and mid-level cues (LMLC) [1]

- Dense and sparse reconstruction (DSR) [17]

- Graph-based manifold ranking (GMR) [12]

- Minimum barrier (MB+) [2]

- Region based contrast (RC) [3]

- Salient region detection via high dimension color transform (HDCT) [8]

- Superpixel based saliency (SP) [14]

- Robust background detection (RBD) [20]

- Multiscale Deep Features (MDF) [23]

- Deep Contrast Learning (DCL) [26]

5.1   Dataset description

In our experiments, five publicly available datasets including **MSRA10K**, **DUTOMRON**, **THUR15K**, **ECSSD** and **PASCAL-S** are employed for performance assessment.

The MSRA10K dataset [3] contains 10000 images with pixel-level salient object labeling [35, 76]. This database has various image categories such as animals, flowers, humans and natural scenes, et al, where most images contain only one salient object. Given the large size and wide variety of contents, this dataset is very challenging for testing the efficacy of relevant saliency detection approaches. The THUR15K dataset [77] consists of 15000 images, which are divided into five categories: butterfly, coffee mug, dog jump, giraffe, and plane. Since it does not contain a salient region labeled for every image, we only use those labeled images (6232 in total) in our experiment for testing. The DUTOMRON dataset [12] is a very challenging database that contains 5166 images. Each image has one or more saliency objects and complex background. The ground truth is labeled by several experienced participants who are familiar with the goal of saliency detection.

The ECSSD dataset [78] and the PASCAL-S dataset [79] are two other challenging databases. ECSSD has 1000

semantically meaningful images with the complicated background, which is also widely used for saliency detection [35]. PASCAL-S has 850 natural images and contains people, animal, vehicles, and indoor objects. This dataset is widely used to recognize an object from a number of visual object classes in the real world [80].

5.2   Evaluation criteria

For quantitative performance assessment of the proposed saliency detection algorithm, several commonly used metrics are adopted in our experiments, which include the precision-recall curve (PR), average precision (AP), receiver operating characteristics (ROC) curve and area under the ROC curve (AUC). By varying a threshold from 1 to 255 and applying it to the determined saliency map, a series of binary images indicating the detected saliency objects can be produced, from which the PR, ROC curve and AUC can be obtained for quantitative assessment.

The PR curve is formed by the true positive rate (TPR, also namely recall) versus positive predictive value (PPV, also namely precision) and the ROC curve is formed by the false positive rate (FPR) versus TPR. The three rates including TPR, PPV and FPR are determined by $TPR = \frac{T_p}{T_p + F_n}, PPV = \frac{T_p}{T_p + F_p}, FPR = \frac{F_p}{T_n + F_p}$, where T_p, F_p, T_n and F_n respectively refer to the number of correctly detected foreground pixels of the salient object, incorrectly detected foreground pixels (false alarms), correctly detected background pixels (non-objects) and incorrectly detected background pixels (or missing pixels from the object). Specifically, these four numbers can be calculated by comparing the binary masks of the detected image and the ground truth.

In addition, the F-measure defined below is also used for comprehensive performance assessment:

$$F_{measure} = \frac{(1 + \beta) \cdot Precision \cdot Recall}{\beta \cdot Precision + Recall} \tag{24}$$

where the parameter β is set to 0.3 to combine the precision and the recall rate as suggested in [18].

5.3   Assessment of the obtained saliency maps

To evaluate the performance of the proposed GLGOV method, we show comprehensive comparison results using the PR curve and AUC values on all the five datasets. For subjective assessment, several typical examples with either large objects or complicated backgrounds are shown in Fig. 4 for comparison. As can be seen, most of these benchmarking methods fail to highlight the objects as a whole or with a high contrast. However, our proposed method can successfully suppress the background regions and maintain the boundaries for the salient object due to the gestalt

| Image | GT | GLGOV | RC[3] | MB+[2] | HDCT[8] | LMLC[1] | DSR[17] | GMR[12] | RBD[20] | SP[14] | MDF[23] |
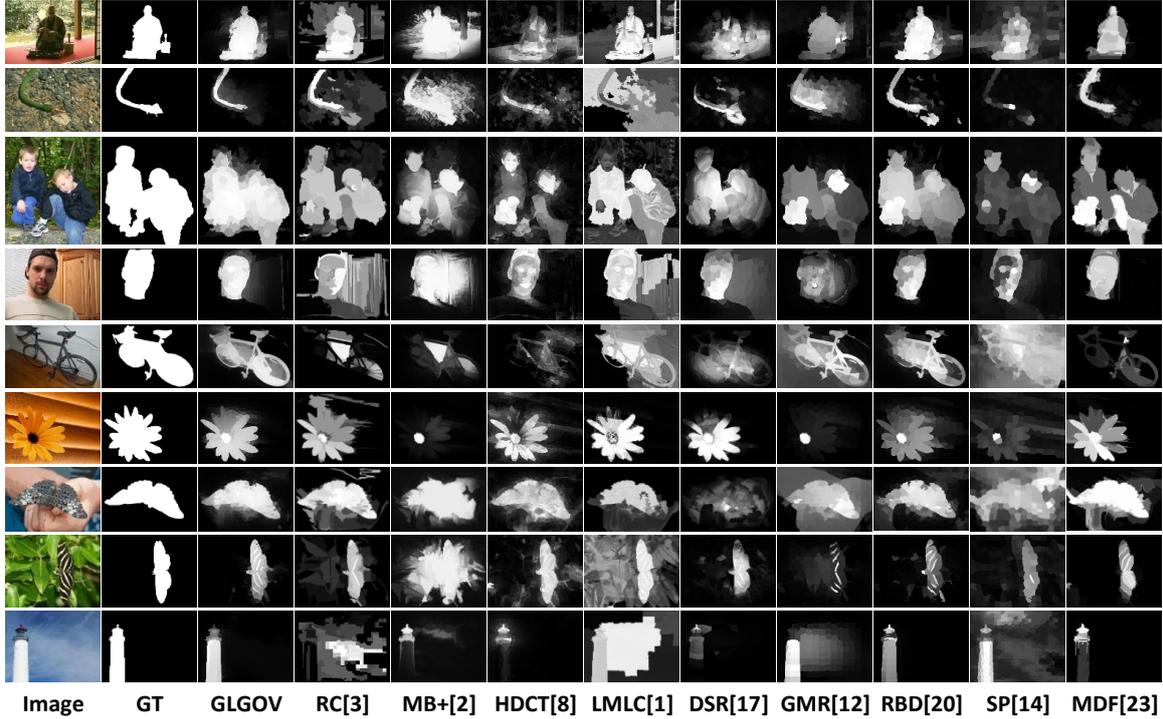
Fig. 4. Visual comparison. The ground truth (GT) in shown in the second column.

law guided cognitive framework. In addition, since the object can be well highlighted, this can further facilitate some potential applications e.g. classification of butterflies (rows 7 and 8) and flowers (row 6) and recognition of people (rows 3 and 4).

For quantitative assessment, the results are compared using AP, AUC measurement and running time for each test image in Table II. In total there are 10 approaches benchmarked with ours in Table II, where the first eight are unsupervised, and the last two are supervised ones using deep learning. All the approaches are tested on a computer with Intel Dual Core i5-4210U 1.7 GHz CPU and 4GB RAM, where for consistency GPU is absent for deep learning based approaches. Due to hardware configuration reasons, we cannot implement DCL hence we use their published saliency maps which are only available for two datasets, i.e. ECSSD and DUTOMRON.

For unsupervised approaches, the proposed approach yields the highest AUC in all the five datasets, followed by DSR, though the AP value from our approach is slightly less than those of DSR except the MSRA10K dataset. However, our approach is 3.5 times faster than DSR. This has validated the efficacy of the proposed approach, especially the gestalt laws and background connectivity used in guiding the process of saliency detection. Although MB+ is the third or fourth place in this group of experiments, the extremely high efficiency makes it a good candidate

20

Table II

AUC, AP and running time measurement (top two unsupervised methods are highlighted in red and green). *: deep learning based method

Method	Time (s)	MSRA10K		PASCAL-S		ECSSD		DUTOMRON		THUR15K		Overall(OA)		Average(AG)	
		AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP	AUC	AP
GLGOV	1.68	0.967	0.888	0.868	0.680	0.915	0.773	0.902	0.539	0.905	0.580	0.930	0.715	0.911	0.692
MB+ [2]	0.05	0.955	0.835	0.859	0.633	0.906	0.695	0.893	0.484	0.900	0.542	0.921	0.665	0.903	0.638
RC [3]	0.25	0.936	0.838	0.707	0.348	0.893	0.733	0.893	0.503	0.896	0.568	0.905	0.669	0.865	0.598
LMLC [1]	140	0.936	0.721	0.793	0.516	0.850	0.570	0.818	0.374	0.853	0.447	0.878	0.557	0.850	0.526
HDCT [8]	4.02	0.941	0.784	0.807	0.628	0.868	0.710	0.867	0.506	0.878	0.541	0.899	0.648	0.872	0.634
DSR [17]	5.85	0.959	0.878	0.866	0.699	0.915	0.788	0.900	0.578	0.902	0.612	0.925	0.729	0.908	0.711
GMR [12]	0.5	0.954	0.882	0.860	0.680	0.894	0.748	0.894	0.544	0.886	0.57	0.893	0.706	0.873	0.681
RBD [20]	0.25	0.944	0.876	0.822	0.663	0.890	0.763	0.854	0.526	0.856	0.579	0.874	0.643	0.846	0.612
SP [14]	1.2	0.923	0.810	0.780	0.576	0.848	0.682	0.837	0.475	0.843	0.518	0.917	0.701	0.898	0.644
MDF* [23]	200	0.973	0.871	0.908	0.762	0.941	0.829	0.917	0.649	0.921	0.611	0.943	0.746	0.932	0.745
DCL* [26]	75	-	-	-	-	0.971	0.897	0.934	0.675	-	-	0.940	0.711	0.952	0.786

for particular applications, i.e. online processing. It is worth noting that although RC, GMR and RBD have very low running time, the AUC and AP measurements they achieve on these datasets seem inferior. For RC, the reason for the degraded performance is mainly due to the hard constraints it used to reduce the saliency value near the image borders. These seem to work fine in the MSRA10K dataset. However, given a large amount of images along with large variations in terms of image contents and complex background such as PASCAL-S, these constraints become less effective in refining the detected saliency maps. Similar to GMR, it works well on MSRA 10K and ECSSD, but fails to process complicated images such as THUR15K and DUTOMRON. For RBD, due to the lack of effective foreground detection, its precision isn't good enough which also leads to its inferior segmentation performance (Table IV). For HDCT, the high dimensional color transform used increases the success of foreground and background separation but increase the running time as well. Moreover, as aforementioned, this method does not totally fit the HVS and results in undesirable results when both the foreground and background contain the same color elements. For LMLC and SP, regardless of the running time, their performance seems to be quite low due to the lack of effective post-processing to further refine the detected saliency maps. Thanks to our cognitive framework, the way our proposed method detects the salient object fit human visual attention very well. This has helped us suppress the background and highlight the main objects more effectively. Therefore, the results of several datasets show much less inconsistency. It is believed that our computation cost can be significantly improved by transplanting the code from

MATLAB to C++ implementation.
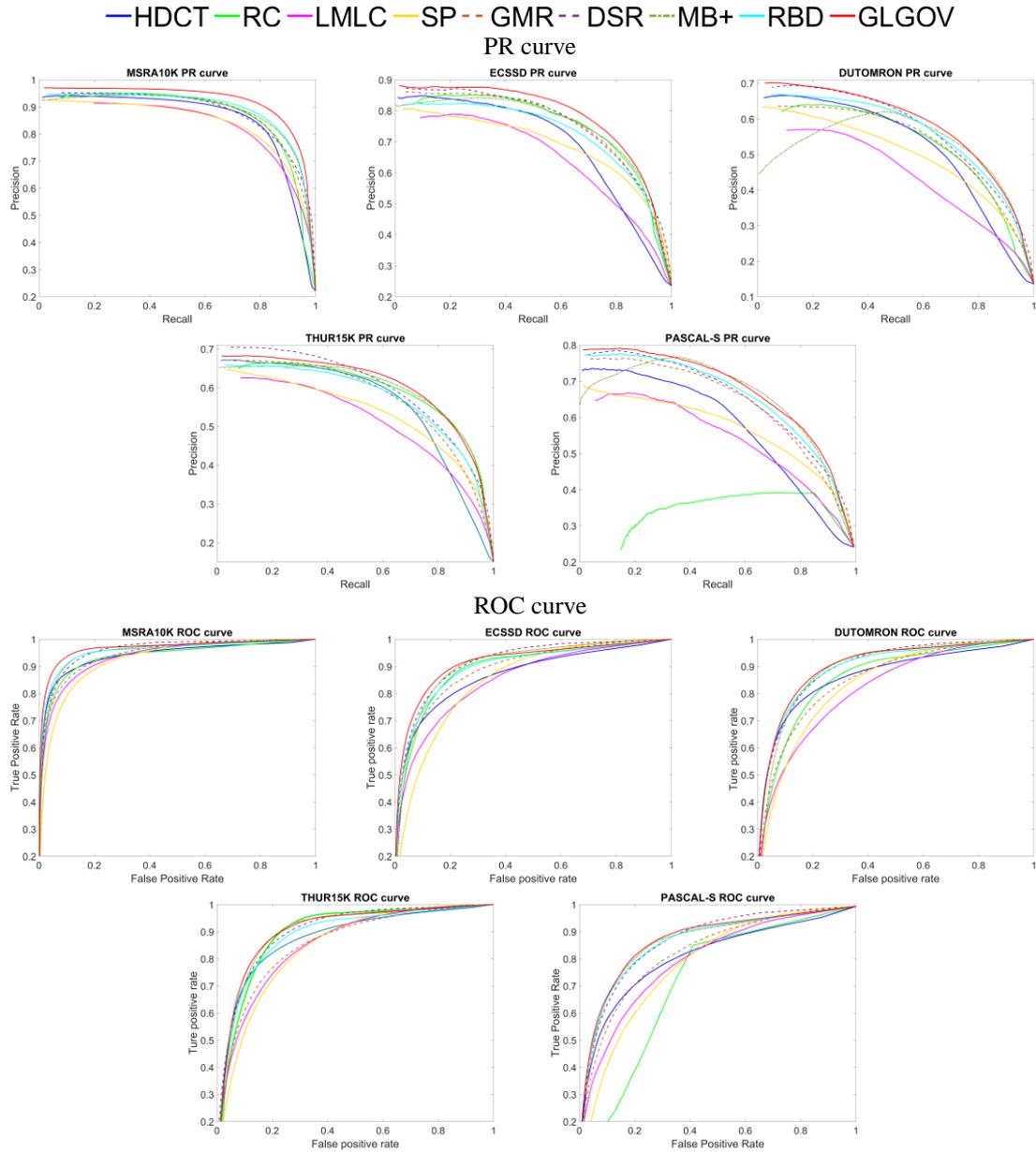
Fig. 5. PR and ROC curve.

For deep learning based supervised approaches including MDF and DCL, unsurprisingly they produce better results in terms of higher AUC and AP values than unsupervised ones, however, they suffer from lengthy training and testing time, also their results seem sensitive to the learning strategies used. This has constrained their applicability for specific tasks that needs nearly real-time response. For MDF, it gains 2% in AUC and 3.5% in AP than unsupervised approach, possibly due to the introduced multiscale CNN deep learning. Nevertheless, for the challenging MSRA10K

dataset, MDF gains 0.6% in AUC but loses 1.6% in AP when comparing to our proposed unsupervised approach, despite of the fact it uses 25% of the samples i.e. 2500 images for training. This again shows the potential limitation of the supervised approach where unsupervised approach can supplement.

To further evaluate the performance of these approaches, we plot in Fig. 5 the PR curves and the ROC curves for the results obtained from the five datasets. For better visual effect, we only compare in Fig. 5 the results from the unsupervised approaches, as the advantages of deep learning based supervised methods have been discussed according to the results in Table II. As seen in Fig. 5, our approach almost outperforms all other unsupervised methods, especially on the MSRA10K and ECSSD datasets, yet the performance on the rest three datasets appears quite comparable to DSR. Although the curves from MB+ are close to those from our GLGOV approach, the much lower AP as shown in Table II indicates more false alarms in the detected results.

5.4 Key component analysis

In this subsection, we discuss the effect of several key components in the proposed method, where all the evaluations are carried out on the MSRA 10k dataset due to its popularity. As the proposed GLGOV framework is actually a multi-stage approach, in the following we assess the contributions of three major components of our algorithm, which include the feature contrast map *FC*, initial saliency map after the first refinement *ISA*, and the final saliency map with the second refinement *FSA*.
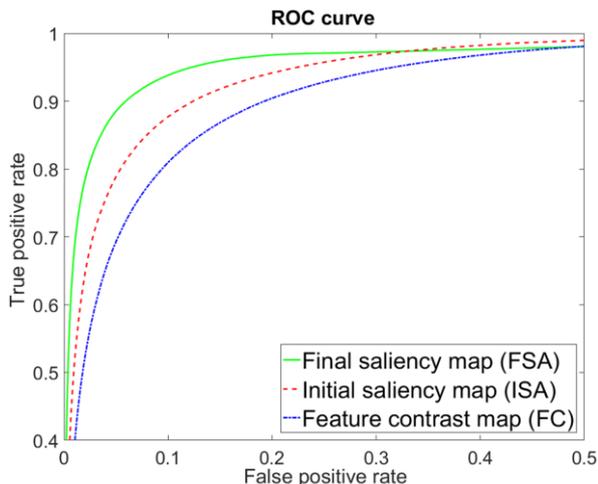
Fig. 6. Results of GLGOV with various settings on the MSRA10K dataset.

Table III
AUC values and running time for our GLGOV approach
under various settings on the MSRA 10K dataset.

Components	FC	ISA	FSA
AUC	0.9330	0.9558	0.9667
Time (s/image)	1.16	1.47	1.68

Fig. 6 shows the ROC curves obtained from these three settings, where the AUC measurements are given and compared in Table III. For the feature contrast map *FC*, the result with an AUC at 93.3% seems undesirable. After applying the figure and ground in the first-stage refinement, the AUC reaches 95.58% with an increase of 2.28%. By further adding background connectivity model for the second-stage refinement, the AUC becomes 96.67%, i.e. an additional gain of 1.09%. Meanwhile, the running time has been increased from 1.16s to 1.47s and 1.68s after

introducing the first and second stage refinement, respectively. This has clearly demonstrates the contribution of the key components in our proposed GLGOV framework.

5.5   Validation on image segmentation

Based on the determined saliency maps, the salient objects can be extracted as binary masks, which can be further applied for performance assessment of the saliency detection approaches. Herein the OTSU approach [60] is used for adaptive thresholding to generate the binary masks of salient objects. For quantitative performance assessment, we calculate the average F-measure of all the test images over their ground truth maps and report the results in Table IV. As can be seen, our proposed GLGOV model consistently produces the highest F-measure on the MSRA10K, PASCAL-S and the ECSSD datasets in comparison to other unsupervised peers, and also the second best on the DUTOMRON and THUR15K datasets after DSR. If taking the five datasets as a whole, our proposed approach outperforms the second best, DSR, 1.5% and 1.1% in terms of the overall and average F-measure, respectively.

It is worth noting that the deep learning based supervised approaches are unsurprisingly high and surpass all unsupervised ones. Nevertheless, there is still some space for further improvement of their learning strategies. For example, the F-measure from MDF is slightly less than our approach on the MSRA10K dataset, which shows that training on 2500 images seems insufficient to fully learn the characteristics of 10k images. This drawback is possibly overcome in DCL, as DCL seems more effective than MDF in the tested two datasets, ECSSD and DUTOMRON. Again, it shows that the performance of deep learning based approaches can be very sensitive to the learning strategy

Table IV
F-measure of segmented results (top two unsupervised methods are highlighted in red and green). *: deep learning based method

Method	MSRA10K	PASCAL-S	ECSSD	DUTOMRON	THUR15K	Overall(OA)	Average(AG)
Proposed	0.8810	0.6625	0.7592	0.6041	0.6042	0.7320	0.7022
RC [3]	0.8395	0.4191	0.7277	0.5647	0.5947	0.6926	0.6291
HDCT [8]	0.8348	0.6074	0.7105	0.5947	0.5881	0.7017	0.6671
LMLC [1]	0.7501	0.5328	0.5880	0.4340	0.4816	0.5930	0.5573
SP [14]	0.8106	0.5689	0.6737	0.5367	0.5456	0.6640	0.6271
GMR [12]	0.8521	0.6467	0.7425	0.5959	0.5924	0.7133	0.6859
DSR [17]	0.8374	0.6476	0.7388	0.6205	0.6114	0.7174	0.6911
MB+ [2]	0.8484	0.6614	0.7226	0.5939	0.5977	0.7124	0.6848
RBD [20]	0.8610	0.654	0.7178	0.6011	0.5851	0.7156	0.6838
MDF* [23]	0.8794	0.7346	0.8173	0.6948	0.6429	0.7670	0.7538
DCL* [26]	-	-	0.8968	0.7375	-	0.7633	0.8172

used, regardless the extremely high computational resources and computational cost needed.

6   CONCLUSIONS

Inspired by both Gestalt laws optimization and background connectivity theory, in this paper, we proposed GLGOV as a cognitive framework to combine bottom-up and top-down vision mechanisms for unsupervised saliency detection. Experimental results over five publicly available datasets have shown that our method helps to produce the best overall accuracy and average accuracy when benchmarking with a number of state-of-the-art unsupervised techniques. Additional assessments in terms of the PR curve, ROC curve, F-measure, AUC and AP have also verified the efficacy of the proposed approach.

The most important finding in this paper is the efficacy of bottom-up and top-down mechanisms for saliency detection, which are actually guided by necrologies such as Gestalt laws and background connectivity. On the one hand, the aim of saliency detection is to enable computers to recognize the salient object like human. On the other hand, Gestalt laws are the main theories that describe the mechanism of HVS, whilst background connectivity can reflect our visual cortex reaction to stimuli. As such, these necrologies can be well introduced into the process of and support the modelling of saliency detection. Our outcomes showed that with the guidance of necrologies, the proposed unsupervised saliency methodology consistently produces good results on different datasets. Although there is still some gap to deep learning based supervised approaches, unsupervised approach may supplement in cases there are no sufficient training samples and/or with limited computational resources.

For future work, we will focus on more in-depth guidance from Gestalt laws on saliency detection, where the laws of closure and continuity can be injected to further improve the performance. Texture feature and deep learning models will also be considered for saliency detection beyond color contrast where semi-supervised or weakly supervised learning can be explored.

REFERENCES

[1]   Y. Xie, H. Lu, and M.-H. Yang, "Bayesian saliency via low and mid level cues," *IEEE Transactions on Image Processing,* vol. 22, pp. 1689-1698, 2013.

[2]   J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 fps," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1404-1412.

[3]   M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 37, pp. 569-582, 2015.

[4]   L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis And Machine Intelligence,* vol. 20, pp. 1254-1259, Nov 1998.

[5]   J. M. Wolfe, "Guided Search 2.0 A revised model of visual search," *Psychonomic Bulletin & Review,* vol. 1, pp. 202-238, 1994.

[6]   C. Koch and S. Ullman, "Shifts in selective visual attention: towards the underlying neural circuitry," in *Matters of intelligence*, ed: Springer, 1987, pp. 115-141.

[7]   R. Desimone and J. Duncan, "Neural mechanisms of selective visual attention," *Annual review of neuroscience,* vol. 18, pp. 193-222, 1995.

[8]   J. Kim, D. Han, Y. W. Tai, and J. Kim, "Salient Region Detection via High-Dimensional Color Transform," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014.

[9]   Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *Proceedings of the eleventh ACM international conference on Multimedia*, ed: ACM, 2003, pp. 374-381.

[10]   J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," presented at the Advances in neural information processing systems, 2006.

[11]   F. Katsuki and C. Constantinidis, "Bottom-up and top-down attention: Different processes and overlapping neural systems," *The Neuroscientist,* vol. 20, pp. 509-521, 2014.

[12]   C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2013, pp. 3166-3173.

[13]   X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2007.

[14]   Z. Liu, L. Meur, and S. Luo, "Superpixel-based saliency detection," presented at the 14th International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS), 2013.

[15]   C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: Bottom-up versus top-down," *Current Biology,* vol. 14, pp. R850-R852, 2004.

[16]   R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *Computer Vision Systems*, ed: Springer, 2008, pp. 66-75.

[17]   X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 2976-2983.

[18]   R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," presented at the IEEE conference on Computer vision and pattern recognition (CVPR), 2009.

[19]   R. Achanta and S. Süsstrunk, "Saliency detection using maximum symmetric surround," presented at the 17th IEEE International Conference on Image Processing (ICIP) 2010.

[20]   W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 2814-2821.

[21]   E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *European Conference on Computer Vision*, ed: Springer, 2010, pp. 366-379.

[22]   K. R. Gegenfurtner, "Cortical mechanisms of colour vision," *Nat Rev Neurosci,* vol. 4, pp. 563-72, Jul 2003.

[23]   G. Li and Y. Yu, "Visual saliency detection based on multiscale deep CNN features," *IEEE Transactions on Image Processing,* vol. 25, pp. 5012-5024, 2016.

[24]   M. M. Cheng, G. X. Zhang, N. J. Mitra, X. Huang, and S. M. Hu, "Global contrast based salient region detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011.

[25]   E. Doi, J. L. Gauthier, G. D. Field, J. Shlens, A. Sher, M. Greschner, *et al.*, "Efficient coding of spatial information in the primate retina," *The Journal of Neuroscience,* vol. 32, pp. 16256-16264, 2012.

[26]   G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 478-487.

[27]   N. R. Carlson, H. Miller, J. W. Donahoe, and G. N. Martin, *Psychology: The Science of Behavior*. Ontario, CA: Pearson Education Canada, 2010.

[28]   N. Al-Aidroos, C. P. Said, and N. B. Turk-Browne, "Top-down attention switches coupling between low-level and high-level areas of human visual cortex," *Proceedings of the National Academy of Sciences,* vol. 109, pp. 14675-14680, 2012.

[29]   W. Zou and N. Komodakis, "HARF: Hierarchy-associated rich features for salient object detection," in *Proceedings of the IEEE International Conference on Computer Vision*, 2016, pp. 406-414.

[30]   Y. Tang and X. Wu, "Saliency detection via combining region-level and pixel-level predictions with CNNS," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* vol. 9912 LNCS, ed, 2016, pp. 809-825.

[31] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5455-5463.

[32] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep image saliency computing via progressive representation learning," *IEEE transactions on neural networks and learning systems,* vol. 27, pp. 1135-1149, 2016.

[33] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3183-3192.

[34] J. J. DiCarlo, D. Zoccolan, and N. C. Rust, "How does the brain solve visual object recognition?," *Neuron,* vol. 73, pp. 415-434, 2012.

[35] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Transactions on Image Processing,* vol. 24, pp. 5706-5722, 2015.

[36] A. Aksac, T. Ozyer, and R. Alhajj, "Complex networks driven salient region detection based on superpixel segmentation," *Pattern Recognition,* vol. 66, pp. 268-279, 2017.

[37] M. Xu, L. Jiang, Z. Ye, and Z. Wang, "Bottom-up saliency detection with sparse representation of learnt texture atoms," *Pattern Recognition,* vol. 60, pp. 348-360, 2016.

[38] X. Sun, Z. He, C. Xu, X. Zhang, W. Zou, and G. Baciu, "Diversity induced matrix decomposition model for salient object detection," *Pattern Recognition,* vol. 66, pp. 253-267, 2017.

[39] M. Iqbal, S. S. Naqvi, W. N. Browne, C. Hollitt, and M. Zhang, "Learning feature fusion strategies for various image types to detect salient objects," *Pattern Recognition,* vol. 60, pp. 106-120, 2016.

[40] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280-287.

[41] A. Furnari, G. M. Farinella, and S. Battiato, "An Experimental Analysis of Saliency Detection with Respect to Three Saliency Levels," in *ECCV Workshops (3)*, 2014, pp. 806-821.

[42] H. Fu, Z. Chi, and D. Feng, "Attention-driven image interpretation with application to image retrieval," *Pattern Recognition,* vol. 39, pp. 1604-1621, 2006.

[43] S.-M. Hu, T. Chen, K. Xu, M.-M. Cheng, and R. R. Martin, "Internet visual media processing: a survey with graphics and vision applications," *The Visual Computer,* vol. 29, pp. 393-405, 2013.

[44] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, and X. Wu, "Visual-textual joint relevance learning for tag-based social image search," *IEEE Transactions on Image Processing,* vol. 22, pp. 363-376, 2013.

[45] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE Transactions on Image Processing,* vol. 19, pp. 185-198, 2010.

[46] Z. Li, S. Qin, and L. Itti, "Visual attention guided bit allocation in video compression," *Image and Vision Computing,* vol. 29, pp. 1-14, 2011.

[47] C. Christopoulos, A. Skodras, and T. Ebrahimi, "The JPEG2000 still image coding system: an overview," *IEEE Transactions on Consumer Electronics,* vol. 46, pp. 1103-1127, 2000.

[48] A. Ninassi, O. L. Meur, P. L. Callet, and D. Barbba, "Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric," presented at the International Conference on Image Processing (ICIP) 2007.

[49] H. Liu and I. Heynderickx, "Visual attention in objective image quality assessment: based on eye-tracking data," *IEEE Transactions on Circuits and Systems for Video Technology,* vol. 21, pp. 971-982, 2011.

[50] Q. Ma and L. Zhang, "Saliency-based image quality assessment criterion," in *Advanced Intelligent Computing Theories and Applications. With Aspects of Theoretical and Methodological Issues*, ed: Springer, 2008, pp. 1124-1133.

[51] Z. Liu, R. Shi, L. Shen, Y. Xue, K. N. Ngan, and Z. Zhang, "Unsupervised salient object segmentation based on kernel density estimation and two-phase graph cut," *IEEE Transactions on Multimedia,* vol. 14, pp. 1275-1289, 2012.

[52] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," presented at the IEEE 12th International Conference onComputer Vision, 2009.

[53] Z. Tu, Z. Guo, W. Xie, M. Yan, R. C. Veltkamp, B. Li*, et al.*, "Fusing disparate object signatures for salient object detection in video," *Pattern Recognition,* vol. 72, pp. 285-299, 2017.

[54] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, pp. 1915-1926, 2012.

[55] J. Han, S. He, X. Qian, D. Wang, L. Guo, and T. Liu, "An object-oriented visual saliency detection framework based on sparse coding representations," *IEEE transactions on circuits and systems for video technology,* vol. 23, pp. 2009-2021, 2013.

[56] J. Han, D. Zhang, G. Cheng, L. Guo, and J. Ren, "Object detection in optical remote sensing images based on weakly supervised learning and high-level feature learning," *IEEE Transactions on Geoscience and Remote Sensing,* vol. 53, pp. 3325-3337, 2015.

[57] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology,* vol. 12, pp. 97-136, 1980.

[58] G. Klir and B. Yuan, *Fuzzy sets and fuzzy logic* vol. 4: Prentice hall New Jersey, 1995.

[59] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," presented at the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012.

[60] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica,* vol. 11, pp. 23-27, 1975.

[61] M. Wertheimer, "Untersuchungen zur Lehre von der Gestalt. II," *Psychological Research,* vol. 4, pp. 301-350, 1923.

[62] M. Wertheimer, "Laws of organization in perceptual forms," 1938.

[63] A. Desolneux, L. Moisan, and J.-M. Morel, "Computational gestalts and perception thresholds," *Journal of Physiology-Paris,* vol. 97, pp. 311-324, 2003.

[64] X. Li and G. D. Logan, "Object-based attention in Chinese readers of Chinese words: Beyond Gestalt principles," *Psychonomic bulletin & review,* vol. 15, pp. 945-949, 2008.

[65] G. Kootstra and D. Kragic, "Fast and bottom-up object detection, segmentation, and evaluation using Gestalt principles," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, 2011, pp. 3423-3428.

[66] Z. Wang and B. Li, "A two-stage approach to saliency detection in images," presented at the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2008.

[67] G. Kootstra, N. Bergström, and D. Kragic, "Gestalt principles for attention and segmentation in natural and artificial vision systems," presented at the ICRA 2011 Workshop on Semantic Perception, Mapping and Exploration (SPME), Shanghai, China, 2011.

[68] J. Wu and L. Zhang, "Gestalt saliency: Salient region detection based on gestalt principles," presented at the 20th IEEE International Conference on Image Processing (ICIP), 2013.

[69] Z. Ren, Y. Hu, L.-T. Chia, and D. Rajan, "Improved saliency detection based on superpixel clustering and saliency propagation," presented at the Proceedings of the 18th ACM international conference on Multimedia, 2010.

[70] S. E. Palmer, *Vision science: Photons to phenomenology*: MIT press, 1999.

[71] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 34, pp. 2274-2282, 2012.

[72] J. Van De Weijer, T. Gevers, and A. D. Bagdanov, "Boosting color saliency in image feature detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 28, pp. 150-156, 2006.

[73] N. I. Córdova, A. Tompary, and N. B. Turk-Browne, "Attentional modulation of background connectivity between ventral visual cortex and the medial temporal lobe," *Neurobiology of learning and memory,* vol. 134, pp. 115-122, 2016.

[74] S. Jeannin, L. Cieplinski, J. R. Ohm, and M. Kim, "Mpeg-7 visual part of experimentation model version 9.0," *ISO/IEC JTC1/SC29/WG11 N,* vol. 3914, 2001.

[75] N.-C. Yang, W.-H. Chang, C.-M. Kuo, and T.-H. Li, "A fast MPEG-7 dominant color extraction with new similarity measure for image retrieval," *Journal of Visual Communication and Image Representation,* vol. 19, pp. 92-105, 2008.

[76] T. L. J. Sun, N.-N. Zheng, X. Tang, H.-Y. Shum, and P. Xi'an, "Learning to Detect A Salient Object," presented at the IEEE conference on Computer Vision and Pattern Recognition (CVPR), 2007.

[77] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu, "Salientshape: Group saliency in image collections," *The Visual Computer,* vol. 30, pp. 443-453, 2014.

[78] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1155-1162.

[79] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280-287.

[80] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1265-1274.