



# Saliency Detection Inspired by Topological Perception Theory

Peng Peng<sup>1</sup> · Kai-Fu Yang<sup>1</sup> · Fu-Ya Luo<sup>1</sup> · Yong-Jie Li<sup>1</sup>

Received: 17 September 2020 / Accepted: 7 May 2021 / Published online: 24 May 2021  
© The Author(s), under exclusive licence to Springer Science+Business Media, LLC, part of Springer Nature 2021

## Abstract

The topological perception theory claims that visual perception of a scene begins from topological properties and then exploits local details. Inspired by this theory, we defined the topological descriptor and topological complexity, and we observed, based on statistics, that the saliencies of the regions with higher topological complexities are generally higher than those of regions with lower topological complexities. We then introduced the topological complexity as a saliency prior and proposed a novel unsupervised topo-prior-guided saliency detection system (TOPS). This system is framed as a topological saliency prior (topo-prior)-guided two-level local cue processing (i.e., pixel- and regional-level cues) with a multi-scale strategy, which includes three main modules: (1) a basic computational model of the topological perception theory for extracting topological features from images, (2) a topo-prior calculation method based on the topological features, and (3) a global–local saliency combination framework guided by the topo-prior. Extensive experiments on widely used salient object detection (SOD) datasets demonstrate that our system outperforms the unsupervised state-of-the-art algorithms. In addition, the topo-prior proposed in this work can be used to boost supervised methods including the deep-learning-based ones for fixation prediction and SOD tasks.

**Keywords** Topological perception theory · Topological complexity · Topological saliency prior · Salient object detection · Fixation prediction

## 1 Introduction

The human visual system (HVS) has the astonishing ability to move its attention to the informative regions of a scene rapidly and effortlessly. The mechanism underlying this ability is believed to be useful for human activities as well as computer vision applications, such as object segmentation (Rahtu et al. 2010), image retrieval (Gao et al. 2015), image

compression (Ji et al. 2013), video object tracking (Ma et al. 2017), and scene classification (Borji and Itti 2011).

This fact also raises a fundamental question of “Where visual processing begins” (Chen 2005) in the field of cognitive neuroscience. To answer this question, the school holding the viewpoint of *early feature-analytic: from local to global* claims that “objects are initially decomposed into separable properties and components, and only in subsequent process objects are recognized, on the basis of the extracted features” (Chen 2005).

This idea seems so natural and reasonable: vision begins from simple components and their local geometric properties, such as line segments with slopes, since they are physically simple and computationally easy. This idea of early feature analysis has gained wide acceptance and has almost dominated the current studies of visual cognition. As representative theories, the Feature Integration Theory proposed by Treisman and Gelade (Treisman and Gelade 1980) and the computational approaches to vision by Marr (Marr and David 1982) still have far-reaching implications on current research in computer vision.

---

Communicated by Jiaya Jia.

---

✉ Yong-Jie Li  
liyj@uestc.edu.cn

Peng Peng  
pengpanda.uestc@gmail.com

Kai-Fu Yang  
yangkf@uestc.edu.cn

Fu-Ya Luo  
luofuya@std.uestc.edu.cn

<sup>1</sup> MOE Key Laboratory for Neuroinformation, the School of Life Science and Technology, University of Electronic Science and Technology of China, 610054 Chengdu, China

On the other side, the viewpoint of *early holistic registration* claims that perception processing is *from global to local*: “Wholes are coded prior to perceptual analysis of their separable properties or parts” (Chen 2005). Gestalt psychology is one of the main schools holding this viewpoint (Koffka 2013).

Compared to Gestalt psychology, Topological Perception Theory (TPT), another school holding this idea, claims that the “whole property” is exactly the “topological property” (Chen 1982, 2005) and that topological pattern recognition may be a fundamental function of the visual system (Chen et al. 2003).

As demonstrated in the works on TPT (Chen 1982, 2005; Chen et al. 2003), the perception of global topology occurs prior to the perception of other pattern features, where the “prior” has two strict meanings. “First, it implies that global spatial and temporal organization, determined by topology, are the basis that perception of local geometrical properties depends on; and second, topological perception takes place *earlier* than the perception of local geometrical properties” (Chen 2005). A series of rigorous experiments make TPT seem more natural (Chen et al. 2003; Zhuo et al. 2003; Chen 2005; Wang et al. 2007; He et al. 2015). More discussions about the relation between TPT and Gestalt psychology can be found in (Chen 2005).

The properties preserved under an arbitrary topological transformation are called topological properties (Chen 1982, 2005). A topological transformation is a one-to-one and continuous transformation in topology terminology. Topological properties involve connectivity, the number of holes, and the inside/outside relation (Chen 2005; He et al. 2015). Chen (2005) suggested that the organization principle of surroundedness in figure–ground perception is exactly the topological properties of holes, which implies that the “hole” identifies the object and background at a very early stage in vision. In this work, we only consider the properties that can be extracted from two-dimensional (2D) digital images that contain “holes”. There exist very few methods based on TPT or the topological features aforementioned for object–background segregation tasks, such as salient object detection (SOD).

Consequently, the two main goals of this work are to establish a basic computational model of TPT and apply it to the saliency detection task. Therefore, we build a system inspired by TPT to extract and encode the topological properties (in image processing, these can also be called topological features), and explore the relationship between the topological features and the saliency to accomplish SOD and fixation prediction tasks with real-world images.

To the best of our knowledge, the proposed system is the first attempt to explicitly introduce TPT into saliency detection. Our main contributions include the followings:

- 1) A computational model is proposed for extracting and encoding topological features. To our knowledge, this is the first relatively complete computational TPT model from the basic conceptions to the feasible topological feature extraction scheme employed for real-world images.
- 2) An image topological complexity calculation model is built to demonstrate that topological features can be used for saliency detection. We reveal the close relevance between the topological features and image saliency by conducting statistical analysis on various datasets. Moreover, this conclusion may provide significant cues and ideas for researchers to further explore topological perception theory.
- 3) A topological saliency prior is proposed for saliency detection, and this prior can be used to boost supervised methods including the deep-learning-based ones utilized in fixation prediction and SOD tasks.
- 4) A topological saliency prior-guided saliency detection framework (TOPS) is proposed for combining global–local saliency. This framework follows the core idea of TPT with a two-pathway structure inspired by Guided Search Theory (Wolfe 1994; Wolfe et al. 2011), and can achieve competitive results compared with the unsupervised state-of-the-art methods on SOD tasks.

The rest of this paper is organized as follows. Sect. 2 gives a brief review of saliency detection. In Sect. 3, we introduce our topological saliency detection system. In Sect. 4, we conduct experiments on popular datasets and some extended analyses. Finally, we conclude this work in the last section. The source code and results are available on our lab’s website<sup>1</sup>.

## 2 Related Work

The field of computer vision has witnessed tremendous progress in saliency detection over the past years (Cong et al. 2019; Wang et al. 2019d; Borji 2019; Wang et al. 2019a). There are two main streams of research: human fixation prediction and salient object detection.

Human fixation prediction aims to estimate the regions of interest (ROIs) where human fixation locates in the images. In contrast, SOD tries to detect the attention-grabbing objects in a scene and segment them. Both tasks can be traced back to Feature Integration Theory (Treisman and Gelade 1980) and the concept of *Computational Attention Architecture* proposed by Koch and Ullman (Koch and Ullman 1987). Following Itti’s computational model (Itti et al. 1998; Itti and Koch 2001), hundreds of methods have been proposed to detect saliency from images and videos. A more detailed

<sup>1</sup> <http://www.neuro.uestc.edu.cn/vccl/>

review about eye fixation prediction models can be found in (Borji and Itti 2012; Borji et al. 2012; Zhao and Koch 2013; Borji 2019). Meanwhile, a survey on SOD can be found in (Borji et al. 2014; Cong et al. 2019; Wang et al. 2019a). The close relation between fixation prediction and SOD has been discussed in (Yin et al. 2014). In this section, we will give a brief review of the features adopted by the saliency detection methods.

*Intrinsic feature contrast-based methods:* To achieve the goal of extracting the most conspicuous foreground objects from a scene, many methods use pixel/regional intrinsic feature contrast, including the luminance, color, texture, and depth contrasts. Achanta et al. (Achanta et al. 2009) adopted a frequency-tuned approach to estimate the saliency map by computing the color difference between every pixel and the mean color of an image. Perazzi (Hornung et al. 2012) demonstrated that a regional contrast could be efficiently computed using a Gaussian blurring kernel. Cheng (Cheng et al. 2011) proposed a region-based method by measuring the global contrast between the foreground targets and other regions. Meanwhile, Yan (Yan et al. 2013; Shi et al. 2016) calculated saliency maps by adopting a hierarchical image framework. Some methods also used other feature contrasts like the depth (Peng et al. 2014; Fang et al. 2014; Qu et al. 2017; Song et al. 2017) or pseudo depth (Xiao et al. 2018) and focusness (Jiang et al. 2013b) as a complement to color features.

*Background prior-based methods:* Although the methods based on feature contrast have made great success, they tend to highlight boundaries and neglect the structure of the images. Some methods attempt to introduce some prior information to capture the image structures. One of the most widely used types of prior information is the background prior. For example, Wei et al. (2012) adopted the geodesic distance between the image pixels and the image borders to estimate the background. Zhang and Sclaroff (2013, 2015) proposed an efficient Boolean map-based method (BMS) to estimate the foreground saliency via computing the minimum barrier distance between each pixel and the pixels located on the image borders. Tu et al. (2016) and Huang and Zhang (2017, 2018) further improved the performance by introducing more efficient algorithms.

These kinds of methods sometimes fail when too much of the target objects touch the borders. To tackle this problem, some researchers introduced more robust background estimation algorithms (Zhu et al. 2014; Li et al. 2015a; Gong et al. 2015) or combined the background prior with the intrinsic feature contrast (Qin et al. 2015; Chen et al. 2016; Yuan et al. 2017).

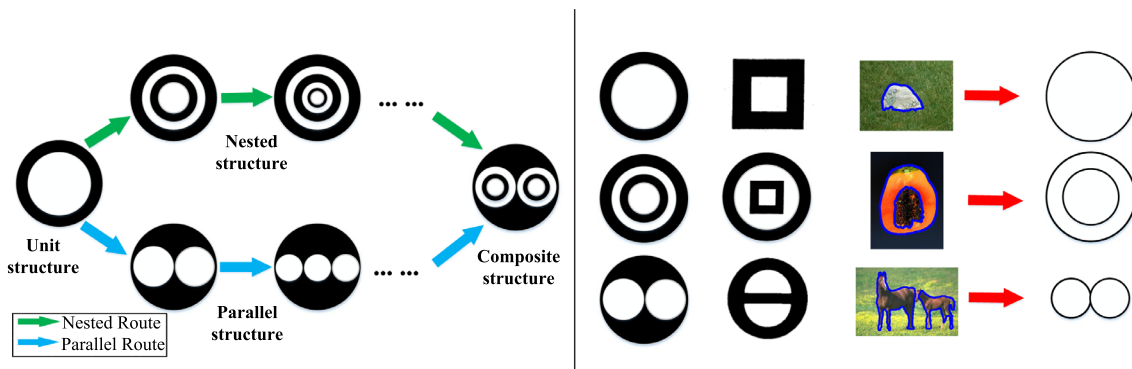
The background prior can be renamed in most conditions by the image *boundary prior*, or, in Gestalt terminology, the *surroundedness*. As TPT claims, “the Gestalt determinant of surroundedness for figure–ground organization is just, in

mathematical language, the topological properties of holes” (Chen 2005). Consequently, the widely used boundary prior is just a special case of topological properties. This concept can also be interpreted by our computational model: the whole image is treated as a pattern where the regions touching the boundaries (i.e., the background) are considered to be the “boundary of a hole,” while the regions that do not touch the boundaries (i.e., the objects) are considered to be the “holes.” Furthermore, a hole is more salient than its boundary when applying the concept of “topological complexity.” More details can be found in Sect. 3.2.

*Other prior-based methods:* Many methods adopt top-down priors other than the background prior to detect the salient subset in images. Many of them introduced information-theoretic knowledge. For instance, Hou and Zhang (2007) developed a spectral residual saliency model based on the assumption that the similarities imply redundancies. Sparse coding and matrix decomposition are also widely adopted (Li et al. 2015b; Peng et al. 2016). Some Bayesian methods follow the guidance of a convex hull of salient points (Xie et al. 2012) or contour-based spatial prior (Yang et al. 2016).

*Learning-based methods:* Recently, a number of learning-based (LB) models (Scharfenberger et al. 2013; Siva et al. 2013; Jiang et al. 2013a), especially deep learning-based (DLB) models (Vig et al. 2014; Zhao et al. 2015; Wang et al. 2016; Kummerer et al. 2017; Kruthiventi et al. 2017; Chen et al. 2017; Wang et al. 2017b; Zhang et al. 2018; Li et al. 2018; Wang et al. 2018; Liu et al. 2018; Zhang et al. 2019; Wang et al. 2019b; He et al. 2019; Zhao and Wu 2019; Zeng et al. 2019; Wu et al. 2019; Wang et al. 2019c) have been proposed. They can usually produce much more precise results than the unsupervised methods on popular datasets. More reviews on deep learning-based methods can be found in (Cong et al. 2019; Borji 2019; Wang et al. 2019a).

Due to the high degree of abstraction and computational complexity of the topological properties, there exist very few computational models for TPT. As a representative work, Huang et al. (2009) proposed a computational topological perceptual organization (CTPO) by establishing topology space under a discrete dot array using a quotient distance histogram. To date, few of the saliency detection methods have directly used the topological properties. Zhang and Sclaroff (2013) computed the color contrast of central pixels to image borders via topological analysis of Boolean maps. In (Gu et al. 2013), the pulse-coupled neural network (PCNN) was employed to produce a “hole filter” to extract the connectivity feature in scenes, which serves as one feature channel. In addition, one of the topological properties, the closure, has been used in (Liu et al. 2017) to detect the closed regions. It has also been regarded as a salient feature and used in (Cheng et al. 2014b) for object proposal.



**Fig. 1** Basic patterns and their topological transformation. Left: categorization of 2D binary patterns and their relationships. Right: simplification of the binary patterns via topological transformation

Moreover, Chen et al. (2019) employed the pieces of saliency results generated from BMS (Zhang and Sclaroff 2015), which were considered as topological features, and a series of center bias maps. Then all of them are integrated into a deep neural network to generate the saliency. Zhou and Gu (2020) proposed to extract topological feature via a pipeline of appearance contrast and segmentation, this feature was used to refine the coarse results generated by a neural network.

Although topological concepts have been adopted in these saliency detection models, what is topological feature is still lacking in clarity and these so-called topological features do not play a major role in their tasks. The relevance between the topological property and the image saliency has not been well demonstrated.

Hence, to address these problems, the system proposed in this work establishes a relatively complete system for topological feature extraction, topological complexity calculation, and the topological complexity-guided global–local saliency combination. The observed positive correlation between the topological features and saliency is the key foundation to the proposed saliency detection system.

### 3 The Proposed Topological Saliency Detection System

We establish in this section a system including the basic computational model for Topological Perception Theory (TPT) and its application for saliency detection. In Sect. 3.1, we build a model to extract topological features from images and represent them as topological descriptors. Then, in Sect. 3.2, a topological complexity calculation method is proposed based on the topological features of an image to reveal the positive correlation between image saliency and the topological features. Moreover, inspired by Guided Search Theory (Wolfe 1994; Wolfe et al. 2011), a topological complexity-guided global–local fusion framework is developed to combine the

topological and local contrast saliency maps; this is presented in Sect. 3.3.

### 3.1 Topological Feature Extraction

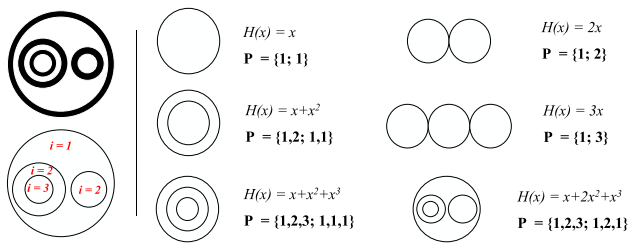
According to TPT (Chen 1982, 2005), the topological property is a global property of the whole pattern, rather than the separated parts. We only consider patterns similar to the ones exhibited in Fig. 1 in this work. These patterns must contain some holes (the enclosed white parts). It is worth clarifying that this does not mean that a target object must have holes in the real world, but it can obtain an enclosed contour extracted by some algorithms. More details about the concept of the “hole” will be discussed in the next subsection, and how we remove the restrictions about “holes” when processing real-world images will be detailed in Sect. 3.1.3.

#### 3.1.1 Categorization and Simplification of Topological Structures

We classify the aforementioned patterns as *unit structure*, *nested structure*, *parallel structure*, and *composite structure*. These patterns are exhibited in Fig. 1 (left).

The *unit structure* is a ring-like pattern that is the most basic structure of our model, and it is used to form the other structures. A pattern with the *nested structure* can be seen as one or more unit structures that are nested together. Similarly, a pattern with the *parallel structure* can be considered as the combination of two or more unit structures in parallel. Meanwhile, a *composite structure* contains an arbitrary number of units, and nested or parallel structures.

“Properties preserved under an arbitrary topological transformation are called topological properties” (Chen 2005). An arbitrary topological transformation means a shape can change significantly to form another shape while maintaining its topological properties. Intuitively, this kind of transformation can be imagined as an arbitrary “rubber-sheet” distortion, in which there is neither break nor fusion (Chen 2005). Since



**Fig. 2** The representation of hole functions and topological descriptors. Left: illustration of the hole function  $H(x) = x + 2x^2 + x^3$  and the region layer labels of a composite structure. Right: hole functions and topological descriptors of different simplified topological structures

we attempt to abstract and encode the topological properties, the local properties, such as scale, orientation, and location, will be ignored.

For easy understanding, we intuitively do a topological transformation to the patterns, which resembles the image skeletonization, and finally obtains the shapes on the right of Fig. 1. As shown in this figure, patterns eventually evolve to some specific shapes that retain the topological properties while ignoring local features. We term these kinds of shapes as *simplified structures*.

It is worth mentioning that the situations where a shape has no hole, for example, a solid shape or dot, will be omitted. Although they are significant for more topological properties, they do not contribute to our saliency detection system. This is because this no-hole shape will finally evolve to a meaningless single point when following the rule of simplification (Fig. 1 (right)). The significance of these no-hole patterns will be discussed in a more sophisticated computational model of TPT in another work.

### 3.1.2 Hole Function and Topological Descriptor

In order to encode the topological features extracted from the simplified structures in a mathematical way, we write a *hole function* as

$$h(x) = \sum_{i=1}^n b_i x^i \quad (1)$$

where  $x$  is the semantic representation of a *hole*; the parameter  $i$  is the layer label of the location of the current hole,  $n$  is the innermost layer label, and the label of the outmost layer is 1.  $b_i$  is the number of holes in the  $i$ -th layer.

*Hole function* is the unique form of topological encoding for a pattern. For example, the hole function of the pattern shown in Fig. 2 (left) is denoted as  $h(x) = x + 2x^2 + x^3$ .

In other words, given the hole function of a pattern, we can know its simplified structure, and vice versa. At the same time, we can also know the labels (i.e., the exponent  $i$ ) of each connected white part (i.e., the *hole*) in a simplified structure.

Although the hole function can encode the topological properties of most 2D patterns following the definition above, the parameter  $x$  is just a semantic representation, unable to be directly used for computation. To address this problem, we introduce a *topological descriptor* derived from the hole function as a vector

$$P = [i; b_i] \quad (2)$$

where  $i = \{1, 2, 3, \dots, n\}$  and  $b_i$  are the same parameters used in Eq. 1. The *topo-descriptor* (topological descriptor) has the same function as the hole function to encode the topological property of a pattern. The main difference between these functions is that the hole function can only demonstrate the topological property of a pattern intuitively while the topo-descriptor can be easily used for computation in our system. An example is illustrated in Fig. 2: the pattern of the left figure contains  $b_i = \{1, 2, 1\}$  different layers according to the layer label  $i = \{1, 2, 3\}$ . Therefore, the topo-descriptor of this pattern is represented by  $P = [1, 2, 3; 1, 2, 1]$ .

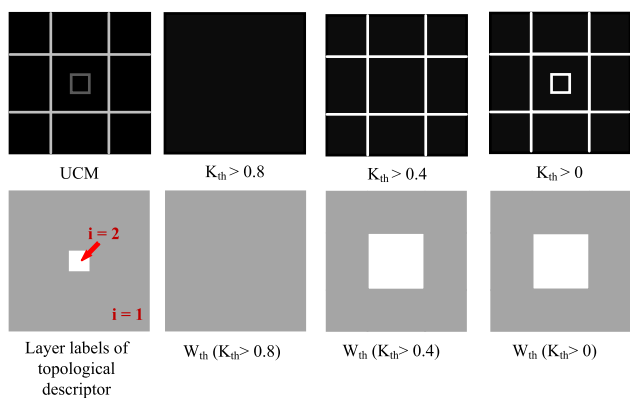
Recently, a study on numerosity perception (He et al. 2015) demonstrates that the connected/enclosed items can lead to robust numerosity underestimation, and the extent of underestimation increases monotonically with the number of connected/enclosed items. This conclusion implies that it is hard to identify the number of inner circles when they are highly nested. Considering this effect, we limit the maximum length of each subvector of the topo-descriptors (that is, we set  $n \leq 5$ ). More topo-descriptor examples can be found in Fig. 2.

Considering the close relation between the hole function and the topo-descriptor, both of them can describe the topological properties of connectedness, the number of holes, and the inside/outside of a pattern. The concept of a hole (i.e.,  $x$ ) is of concern only if the white region is segregated by a circle line. The sum of  $b_i$  (i.e.,  $\sum_{i=1}^n b_i$ ) denotes the total number of holes in a pattern, while the region layer label  $i$  indicates the inside/outside relationship.

### 3.1.3 Extraction from Real-world Images

It is natural to attempt to extract the topological properties from real-world scenes using our computational model. Unfortunately, real-world images are much more complex than the aforementioned patterns. Many researchers believe that image contours contain rich information about the scene context (Zitnick and Dollár 2014; Yang et al. 2016; Liu et al. 2017). Consequently, we use a region-based natural image segmentation method, the ultrametric contour map (UCM) (Arbelaez 2006), to extract the contours of an image; these contours will serve as the simplified structure.

As described in Arbelaez (2006), a UCM is a soft boundary map associated with a family of closed, nested, and non-



**Fig. 3** Illustration of the UCM-defined contours of various confidences, and the corresponding weighting maps. From left to right, first row: a UCM and its binary maps obtained by different thresholds ( $K_{th}$ ); second row: layer labels in the topo-descriptor, and three weighting maps ( $W_{th}$ ) corresponding to the binary maps in the first row. These weighting maps are used to lower the effect of the regions touching the image boundaries when computing the topo-complexity map with Eq. 4

self-intersecting weighted contours, and thus demonstrates a hierarchy of regions that can represent the geometric structures of an image. The different levels of a UCM contain the contours of different confidences, resulting from over-segmentation to under-segmentation. Thresholding between the confidence levels will result in a continuous trade-off between these extremes (Arbelaez 2006). The first row of Fig. 3 presents a simple example of a UCM.

It is generally accepted that adopting the widely used multi-scale strategy to process the coarse-to-fine tasks can achieve better performance (Adelson et al. 1984; Heeger and Bergen 1995). Further, the results of perception organization can be affected by visual acuity, the distance between the subject and the objects, and so on. Consequently, we treat the UCM of an image leveled by different thresholds as the simplified multi-scale structures of this image, which has been introduced in Sect. 3.1.1. On each simplified structure, a hole function or a topo-descriptor can be computed using Eq. 1 or Eq. 2, and then a weighted averaging is applied to the multi-scale topo-descriptors. The first figure of the second row in Fig. 3 illustrates the layer labels of the multi-scale topo-descriptor.

A “hole” in a real-world image is a region defined by a closed contour, and the contour map of an object is treated as the simplified structure. As briefly discussed in Sect. 2, the “surroundedness” in Gestalt terminology is considered as the special cases of the perception of “holes,” in other words, the “surroundedness” is the one-hole situation in our computational model.

The UCM provides us a feasible approach to extract the simplified multi-scale structures of a real-world image. Its properties of closure, nesting, and multi-scale representation ensure the reliability and robustness of our results, which will

not be distorted by small interference, such as edge imperfection.

### 3.2 From Topological Features to Saliency

TPT suggests that “global topological perception is prior to the perception of other features” (Chen 2005). One of the meanings of the “prior” is that “the global spatial and temporal organizations, determined by topology, are the basis that perception of local geometrical properties depends on” (Chen 2005). This suggests that topological features can be used as an important saliency prior to help integrate the local salient cues (such as the color contrast) to achieve better saliency results. Based on this idea, we suppose that there must be an implicit relevance between topological features and visual saliency. Consequently, in order to describe this kind of relevance, we introduce a concept of topological complexity for each region at each scale and an *observation* to bridge the gap between the topological features and the visual saliency. The role of “topological complexity” is to transform the abstract global topological features to intuitive regional saliency density. In the next section, we will propose a topological complexity-guided framework for salient object detection.

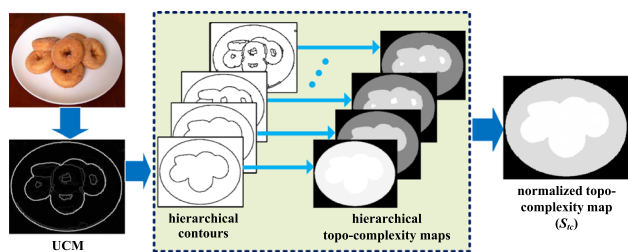
#### 3.2.1 Topological Complexity

The topological complexity function  $F_c$  is computed on a simplified structure described in Sect. 3.1.1. For each region in the structure, we define its topo-complexity as

$$F_c = \sum_{j=1}^n e^{-2(j-1)} \tag{3}$$

where parameter  $j$  is the region’s layer label obtained from the topo-descriptor  $\mathbf{P}$  described by Eq. 2. The whole map of topo-complexity of a single scale structure is denoted as  $F_{map}$ . This expression tells us that the topo-complexity of a region is the cumulative sum of exponential relations on the layer label. Thus, for a region in the map, a larger label corresponds to a higher topo-complexity.

To compute the topo-complexity of each region at each scale of a real-world image, we adopt the UCM used in Sect. 3.1.3. The regions touching the image borders are sometimes considered as the image background (Zhu et al. 2014; Zhang and Sclaroff 2015) or the boundary of holes in our computational model. Generally, the saliency of the background should be lower than that of object regions. Hence, we lower the weight of topo-complexity of the regions touching image borders. The weighting factor  $W_{th}$  is empirically set to a positive value between [0, 1].



**Fig. 4** Flowchart of the hierarchical topo-complexity map calculation

Consequently, we calculate a full topo-complexity map  $S_{ic}$  normalized to  $[0, 1]$  for each real-world image by

$$S_{ic} = MN\left(\frac{1}{N_s} \sum_{n_s=1}^{N_s} W_{th}^{(n_s)} \cdot F_{map}^{(n_s)}\right) \quad (4)$$

where  $MN(\cdot)$  is min–max normalization, the weighting factor  $W_{th}^{(n_s)}$  is a mask at the  $n_s$ -th scale, and the weights of the regions touching the boundaries are 20% of their calculated topo-complexity values.  $N_s$  is the total number of scales created by proper thresholds. Considering that a higher threshold leads to an under-segmentation and we will obtain an incorrect topo-complexity map, we empirically set  $N_s = 4$  with 4 thresholds  $\{0.1, 0.3, 0.5, 0.7\}$  to balance the computational load and the performance in this work. See the examples of  $W_{th}(K_{th} > 0.8)$  in Fig. 3.

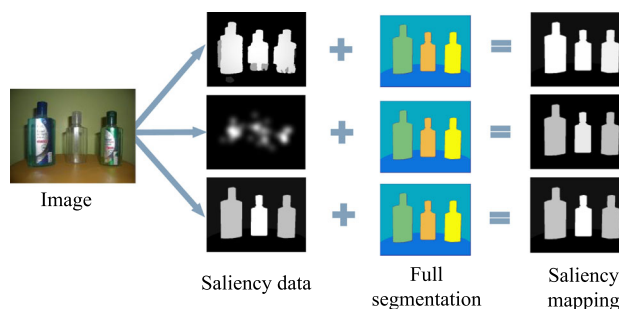
A simple example with three thresholds  $\{0, 0.4, 0.8\}$  is shown in Fig. 3. Here we used these thresholds to get three scales; the last three figures in the second row are the corresponding weighting maps  $W_{th}$ . The computation of topo-complexity map in real-world images is summarized in Fig. 4.

### 3.2.2 An Observation

We observed that, **generally, the saliency of a region with higher topological complexity is higher than that of the regions with lower topological complexities.**

This observation was obtained as follows. We conducted a statistical analysis on the PASCAL-S dataset (Yin et al. 2014). This dataset provides source images, full segmentation of the images, eye fixation data, and human-labeled salient object ground truths. Li’s work on PASCAL-S (Yin et al. 2014) discussed the strong relevance between the tasks of SOD and eye fixation prediction. Similar to Li’s work, here we try to reveal the relevance between the topo-complexity and visual saliency of an object in an image.

The analysis is demonstrated in Fig. 5, and is similar to the operations used in the SBSO dataset’s construction (Yang et al. 2015). The core idea of this process is to compare the similarity of the results of three different saliency density data (i.e., our topo-complexity maps, human fixation maps, and

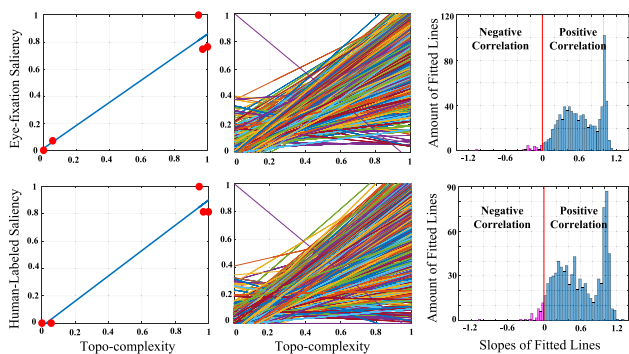


**Fig. 5** Illustration of the saliency mapping for analyzing the relationship between the topo-complexity and the visual saliency. The second column presents three different saliency data (from top down, our topo-complexity, the human eye fixation, and the human-labeled ground truths) are mapped on the full segmentation data to generate new maps. The new maps are presented in the right column, respectively

human-labeled saliency maps). Therefore, corresponding to these different saliency data, three new region-based saliency maps are generated (see the right column in Fig. 5).

Then, we compared every region of saliency density between the newly generated topo-complexity maps with the other two kinds of maps. The results are sketched in Fig. 6, in which the first and second rows show compare topo-complexity with eye fixation and human-labeled saliency, respectively. Let us take one image as an example. The first column is the comparison on the image shown in Fig. 5. In this graph, each region of the full segmentation is denoted as a red point in the graph. The coordinates of red dots are the values of topo-complexity and saliency data (i.e., the eye fixation data and human-labeled saliency data), respectively. Each red dot in Fig. 6 denotes a region. In the full segmentation map in Fig. 5, there are 5 regions with different colors (3 objects and 2 background regions). A fitted line of these points for each image is plotted as a colored line in the graph. A fitted line with a positive slope indicates that the regional topo-complexity has a positive correlation with the density of eye fixation data or the human-labeled saliency. This positive correlation between the topo-complexity and the visual saliency is not by coincidence. We also analyzed such correlations for all 850 of the images in PASCAL-S. These correlations are depicted in the middle columns of Fig. 6. Each fitted line corresponding to one image is drawn in this panel. The right column of this figure shows the statistics of all of the slopes in the form of a histogram. It is clear that the slopes of most fitted lines are positive values (the blue bars) and only a few are negative (the magenta bars).

We obtained similar observations on other saliency datasets, i.e., ImgSal (Li et al. 2013) and Judd (Judd et al. 2009). All these statistics-based observations validate the assumption underlying our model that there is a positive correlation between the topo-complexity and the visual saliency for most scenes.



**Fig. 6** The observed relationship between the topo-complexity and visual saliency. From top to bottom: Topo-complexity with eye fixation and human-labeled saliency. From left to right: the results for one image shown in Fig. 5, results for all 850 of the images from PASCAL-S (Yin et al. 2014), the histogram of the slope values of all fitted lines. In the first column, each red dot denotes the regional correspondence of the saliency mapping between topo-complexity and eye fixation or human-labeled saliency ground truth, and the solid lines represent the linear fittings of these dots. In the middle column, only the fitted lines are drawn, each of which corresponds to one image. Please see the text for details

### 3.2.3 Fixation Prediction

According to the observation described above, we suppose that the extracted topo-complexity map can be used to predict the fixation. To reduce the effect of regional hierarchy and smooth the topo-complexity map, we use a Gaussian filter with specific  $\sigma$  (standard deviation).

Since our  $S_{tc}$  only uses the topo-feature, it is natural to extend it into other fixation prediction methods to promote their performances. We proposed a simple extended strategy for methods such as GBVS (Harel et al. 2007):

$$S_{GBVS*} = MN(\exp(S_{tc_{50}}) + S_{GBVS} + \exp(S_{GBVS})) \quad (5)$$

where  $S_{tc_{50}}$  is the Gaussian-blurred topo-complexity map with  $\sigma = 50$ .

The results in Table 3 show that this strategy is simple and effective. More details about the experiment and the quantitative results are described in Sect. 4.5.

## 3.3 Topological Saliency Prior-guided Framework

### 3.3.1 Two-pathway Combination

The strict meaning of “prior” is written as “global topological perception is prior to the perception of other pattern features” (Chen 2005). Which describes the spatial relations, temporal relations, and causality between the topological properties and the local geometrical properties. We have discussed in the previous section the possibility of treating the topological property as saliency prior to integrating local salient cues.

Several attempts have been made to integrate multiple low-level cues to produce better saliency detection results (Xie et al. 2012; Yang et al. 2016; Lin et al. 2019).

Another famous visual theory, Guided Search Theory (GST) (Wolfe 1994; Wolfe et al. 2011), claims that information processing of the visual search goes along two parallel pathways: a non-selective pathway and a selective pathway. This strategy allows the HVS to rapidly extract global spatial information via the non-selective pathway, which then acts as top-down modulation to guide the processing of local cues in the selective pathway.

Hence, in this section, and inspired by these theories and strategies, a two-pathway SOD framework is proposed and summarized in Fig. 7. In the so-called topological prior-guided saliency detection (TOPS) framework, the topological complexity map is treated as top-down information that can be rapidly obtained, and the local cues are processed in the selective pathway. Extensive analyses in the next section will confirm that our topo-prior guidance performs better than some popular priors such as the CBSP (Yang et al. 2016) and RBD (Zhu et al. 2014).

Electro-physiological evidences of Livingstone (Livingstone and Hubel 1987) and DeYoe (DeYoe and Van Essen 1988) show that the local cues (e.g., color) are coded hierarchically in the visual cortices V1 and V2, and then integrated into higher-level cortices V3 and V4. This tells us that there are many complicated processes in the processing of local features in the brain. Therefore, we design the selective pathway in our framework to mimic the processing of these local cues.

The saliency map with only topo-complexity,  $S_{ots}$ , along the non-selective pathway, is generated by

$$S_{ots} = f(S_{tc} \cdot S_{bias}) \quad (6)$$

The refinement function  $f(x)$  is defined as

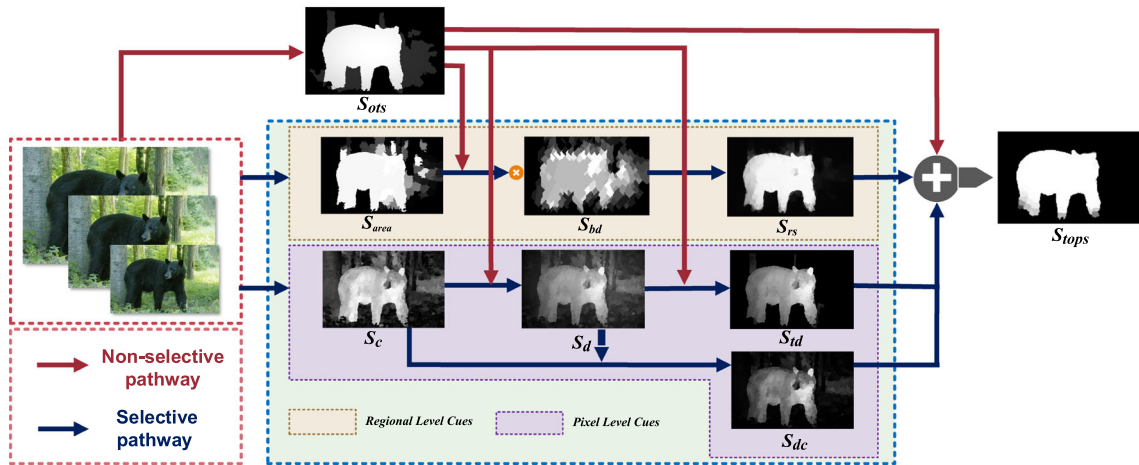
$$f(x) = \frac{1}{1 + e^{-\gamma \cdot (x - \tau)}} \quad (7)$$

where  $\gamma$  and  $\tau$  are empirically set to 10 and 0.5, respectively. The  $S_{bias} = e^{-\|x_k - x_{cen}\|_2^2 / (H/W \cdot \sigma)^2}$  is a center bias with a large  $\sigma$  for an image of size  $W \times H$  (as used in many previous methods (Itti et al. 1998; Yang et al. 2016)). Since Eq.6 demonstrates a topological saliency prior map, we also call  $S_{ots}$  the *topo-prior*.

### 3.3.2 Extraction of Pixel-level Cues

There exist two main processing flows along the selective pathway in Fig. 7, one for pixel-level cues and another for regional-level cues. To compute the pixel-level saliency fea-





**Fig. 7** The proposed framework of topo-prior guided saliency detection (TOPS). This two-pathway-based framework contains the selective and non-selective pathways, which serve to extract the local cues and global prior, respectively. As a kind of global information, our topo-

complexity prior ( $S_{ots}$ ) acts as a global signal to guide the extraction and combination of the local regional and pixel level features along the selective pathway

tures, the color contrast map  $S_c$  is first calculated by

$$S_c(s_k) = \frac{\sum_{\forall s_i \in I} D(s_k, s_i)}{N} \tag{8}$$

where  $D(s_k, s_i)$  is the Euclidean color distance metric between the superpixels  $s_k$  and  $s_i$  in *CIE Lab* color space. All superpixels  $I = \{s_1, s_2, \dots, s_N\}$  are obtained using the toolbox from (Achanta et al. 2012).  $N$  is the total number of superpixels.

However, there still exist two main problems for Eq.8. (1) spatial distances between  $s_k$  and other superpixels  $s_i (i = 1, 2, 3, \dots, N)$  are different, but share the same weights. (2) color contrast between super-pixels share the same weights wherever they are located in background or object areas. To address these problems, the *distance and topological structure-weighted map*  $S_d$  combines a weighting item with two parts for every superpixel according to their location and topo-prior, which is evaluated by

$$S_d(s_k) = \frac{\sum_{\forall s_i \in I} W(s_k, s_i) \cdot D(s_k, s_i)}{N} \tag{9}$$

where

$$W(s_k, s_i) = (1 - S_{ots}(s_i)) \cdot e^{-\frac{|x_i - x_k|^2}{2\sigma^2}} \tag{10}$$

We set  $\sigma = 100$  here.  $x_i$  and  $x_k$  are the center coordinates of the corresponding superpixels  $s_i$  and  $s_k$ . The first part of Eq. 10 uses the reverse of the topo-saliency prior to reduce the saliency of background regions and enhance the saliency of object area. We believe that the contrasts from pixels in background can enhance the saliency of pixels in

the foreground while the contrasts from pixels in foreground can reduce the saliency of pixels in the background. At the same time, the second part emphasizes the spatial distance influence between superpixels by assigning larger weights for the superpixels near  $s_k$  while smaller weights for the superpixels far away from  $s_k$ . Therefore,  $S_d$  computed by Eq. 9 represents a color contrast map modulated by spatial and topological information.

We then multiply  $S_{ots}$  and  $S_d$  to get  $S_{td}$  to simulate the contribution of the topo-complexity in guiding the selective pathway.

$$S_{td}(s_k) = S_{ots}(s_k) \cdot S_d(s_k) \tag{11}$$

Next, we multiply  $S_d$  by  $S_c$  to get  $S_{dc}$  to mimic the inner interaction of local cue processing in the visual cortices.

$$S_{dc}(s_k) = S_d(s_k) \cdot S_c(s_k) \tag{12}$$

After processing Eq.9, Eq.11, and Eq.12, the influence from topo-complexity guidance is penetrated into both  $S_{td}$  and  $S_{dc}$ , and the degree of the influence is  $S_{dc} < S_{td} < S_{ots}$ . Then, the outputs of these two pathways are linearly combined into a *pixel-level saliency map*  $S_{ps}$  as

$$S_{ps}(s_k) = MN(S_{td}(s_k) + S_{dc}(s_k)) \tag{13}$$

where  $MN(\cdot)$  is the min-max normalization.

### 3.3.3 Extraction of Regional-level Cues

In addition to the low-level pixel cues, mid-level regional cues have also been reported as useful for SOD (Liu et al.

2017; Lin et al. 2019). For example, the regions touching image boarders are likely to be the background (Wei et al. 2012; Zhang and Sclaroff 2015). Furthermore, the boundary connectivity of regions have a close relation to saliency and has been proved to be a useful saliency prior in (Zhu et al. 2014). Therefore, the method for calculating the dissimilarity map  $S_{bd}$  between regions and image boarders is adapted from (Wei et al. 2012), and the method for calculating the boundary connectivity map  $S_{bc}$  is adapted from (Zhu et al. 2014). We guided the extraction of these two regional features with our topo-prior to compute the regional-level saliency as

$$S_{rs}(s_k) = MN(\exp(S_{ots}(s_k) + S_{bc}(s_k) + S_{bd}(s_k))) \quad (14)$$

In this way, the topo-complexity guidance is also penetrated into the regional-level saliency.

### 3.3.4 Saliency Fusion and Optimization

The multi-scale strategy is widely used in image processing, and here we adopt it to promote performance. We use 3 scales ( $scl = \{1, 0.75, 0.5\}$ ) of the original image size, and then normalize and combine them.

Finally, an optimization operation is applied after the fusion of pixel-level and regional-level saliency, as shown in Fig. 7, to obtain a more sophisticated saliency map (the *topo-saliency map*)  $S_{tops}$  (Zhu et al. 2014; Liu et al. 2017):

$$S_{tops} = f[opt(S)] \quad (15)$$

where

$$S = MN(\sum_{scl} scl \cdot [S_{ps}^{(scl)} + S_{rs}^{(scl)} + S_{ots}^{(scl)}]) \quad (16)$$

The optimization method  $opt(\cdot)$  is adapted from(Zhu et al. 2014). Let the saliency values of  $N$  superpixels be denoted as  $\{s_i\}_{i=1}^N$ , and the cost function is

$$\sum_{i=1}^N w_i^{(bg)} s_i^2 + \sum_{i=1}^N w_i^{(fg)} (s_i - 1)^2 + \sum_{i,j} w_{ij} (s_i - s_j)^2 \quad (17)$$

We set  $w^{(fg)}$  and  $w^{(bg)}$  as  $S$  and  $1 - S$ , respectively, and the other parameters are set as (Zhu et al. 2014) suggested.

The refinement function  $f(x)$  in Eq.15 is defined by Eq.7, for which we set  $\lambda = 20$  and  $\tau$  to 1.2 times an adaptive threshold computed using Otsu’s binary threshold method (Otsu 1979) to control the overall sharpness.

**Table 1** Parameter Sensitivity Analysis (Color table online)

Param.		ECSSD			PASCAL-S	
$\gamma$	$\tau$	$\sigma$	MAE $\downarrow^a$	Sm. $\uparrow$	MAE	Sm.
10	<b>1.2</b>	<b>100</b>	0.143	<b>0.767</b>	0.187	<b>0.693</b>
<b>20</b>	1.1	<b>100</b>	0.131	<b>0.762</b>	0.177	<b>0.687</b>
<b>20</b>	1.3	<b>100</b>	<b>0.126<sup>b</sup></b>	<b>0.762</b>	<b>0.173</b>	0.684
<b>20</b>	<b>1.2</b>	50	0.131	0.761	0.177	0.684
<b>20</b>	<b>1.2</b>	<b>100</b>	<b>0.127</b>	<b>0.763</b>	<b>0.174</b>	<b>0.685</b>
<b>20</b>	<b>1.2</b>	150	0.131	0.759	0.177	0.683
30	<b>1.2</b>	<b>100</b>	<b>0.125</b>	0.76	<b>0.172</b>	0.682

<sup>a</sup>. The  $\uparrow$  and  $\downarrow$  denote “higher is better” and “lower is better,” respectively.

<sup>b</sup>. **Red, green, and blue** indicate the top three performances, respectively

## 4 Experiments

We conducted experiments on six popular and challenging datasets that covering various scenarios, compared with the state-of-the-art (SOTA) methods. Then, we evaluated the ability for fixation prediction of our topo-complexity prior compared with some leading methods. We also conducted parameter, mechanism, runtime, and ablation analyses to thoroughly evaluate the proposed system. Failure cases are also listed to exhibit the limitations and directions for further improvements.

### 4.1 Parameter Setting

There are some hyper-parameters in this work, which we set either by referring to the literature or setting empirically. For example,  $n \leq 5$  in Eq.1 was inspired by the conclusion of (He et al. 2015); the scale values  $scl = \{1, 0.75, 0.5\}$  in Eq.16 were adopted from (Liu et al. 2017); the parameters in the optimization of Eq.17 were adopted from (Zhu et al. 2014). The scales  $\{0.1, 0.3, 0.5, 0.7\}$  generating topological features in Eq.4,  $\gamma$  and  $\tau$  in Eq.7, Eq. 15 and so on, were set empirically. It should be noted that when setting parameters manually, we tried to balance accuracy and efficiency as much as possible.

Table 1 shows a series of results when changing some of the important parameters, such as the  $\gamma$  and  $\tau$  in Eq. 15 and  $\sigma$  in Eq. 10. The results show that, roughly,  $\tau$  and  $\sigma$  have slight influence to the performance, and the value of  $\gamma$  is inversely proportional to the both values of mean absolute error (MAE) and Structure-measure (Sm.). In other words, larger  $\gamma$  values produce better MAE but worse Sm. Overall speaking, it is appropriate to choose 20, 1.2, and 100 for  $\gamma$ ,  $\tau$  and  $\sigma$ , respectively.

## 4.2 Datasets and Compared Methods for SOD

The proposed methods (the OTS described by Eq. 6 and the TOPS described by Eq. 15) were compared with 15 recently proposed unsupervised state-of-the-art methods on six datasets: MSRA10k (Cheng et al. 2014a), Pascal-S (Yin et al. 2014), HKU-IS (Li and Yu 2015), ECSSD (Shi et al. 2016), DUT-OMRON (Yang et al. 2013), and DUTS-test (Wang et al. 2017b). Unsupervised methods for SOD were compared, including FT (Achanta et al. 2009), GS (Wei et al. 2012), MR (Yang et al. 2013), HC and RC (Cheng et al. 2011), BMS (Zhang and Sclaroff 2015), HS (Shi et al. 2016), RBD (Zhu et al. 2014), RCRR (Li et al. 2015a), CGVS (Yang et al. 2016), SMD (Peng et al. 2016), PDP (Xiao et al. 2018), MST (Tu et al. 2016), WFD (Huang and Zhang 2018), and HCCH (Liu et al. 2017).

In addition, leading supervised methods DRFI (Wang et al. 2017a) and HDCT (Kim et al. 2015), the weakly supervised deep-learning-based (DLB) method WSS (Wang et al. 2017b), and fully supervised DLB methods (RFCN (Wang et al. 2016), PiCANet (Liu et al. 2018), PAGRN (Zhang et al. 2018), RAS (Chen et al. 2018), PFA (Zhao and Wu 2019), PAGE (Wang et al. 2019e), ETF (Zhou and Gu 2020), RASv2 (Chen et al. 2020), and CPD (Wu et al. 2019)) were also compared.

The MSRA10k dataset contains 10k images with relatively simple scenes containing single objects (Cheng et al. 2014a). HKU-IS is also a large-scale image dataset that contains 4447 more challenging and unbiased images (Li and Yu 2015). ECSSD includes 1000 semantically meaningful but structurally complex images (Shi et al. 2016). Meanwhile, PASCAL-S is a very challenging baseline in saliency detection since it usually involves several different objects against a cluttered background (Yin et al. 2014). DUT-OMRON contains 5168 high-quality images with a single object but a more challenging background (Yang et al. 2013). DUTS-Test is the part of DUTS (Wang et al. 2017b) that is used for testing while DUTS-Training is used for training. Both HKU-IS and DUTS are the current most widely used baselines.

## 4.3 Evaluation Metrics on SOD

For a more comprehensive evaluation, we adopted five metrics: precision–recall (PR) curves, F-measure curves, the weighted F-score, the mean absolute error (MAE), and the structure measure (Sm.). Precision and recall were computed by thresholding the saliency map and comparing the binary map with the ground truth. This metric represents the mean precision and recall of all of the saliency maps at different thresholds between [0, 255]. Usually, neither the values of precision nor recall can comprehensively demonstrate the performance of saliency detection algorithms. Hence, the weighted F-score is adopted for a more effective evaluation,

which is defined as

$$F_{\beta}^* = \frac{(1 + \beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \quad (18)$$

considering that the precision is more important than recall (Liu et al. 2010), we set  $\beta^2$  to 0.3 as (Achanta et al. 2009; Cheng et al. 2014a) suggested. The overlap-based metrics introduced above ignore the situation of the correct assignment of non-salient pixels; therefore, the mean absolute error (MAE) score is adopted to address this problem for a more comprehensive comparison (Perazzi et al. 2012). The MAE between the saliency map  $S(x, y)$  with a size of  $W \times H$  pixels and the binary ground truth  $G(x, y)$  is defined as

$$MAE = \frac{1}{W \times H} \cdot \sum_{x=1}^W \sum_{y=1}^H |S(x, y) - G(x, y)| \quad (19)$$

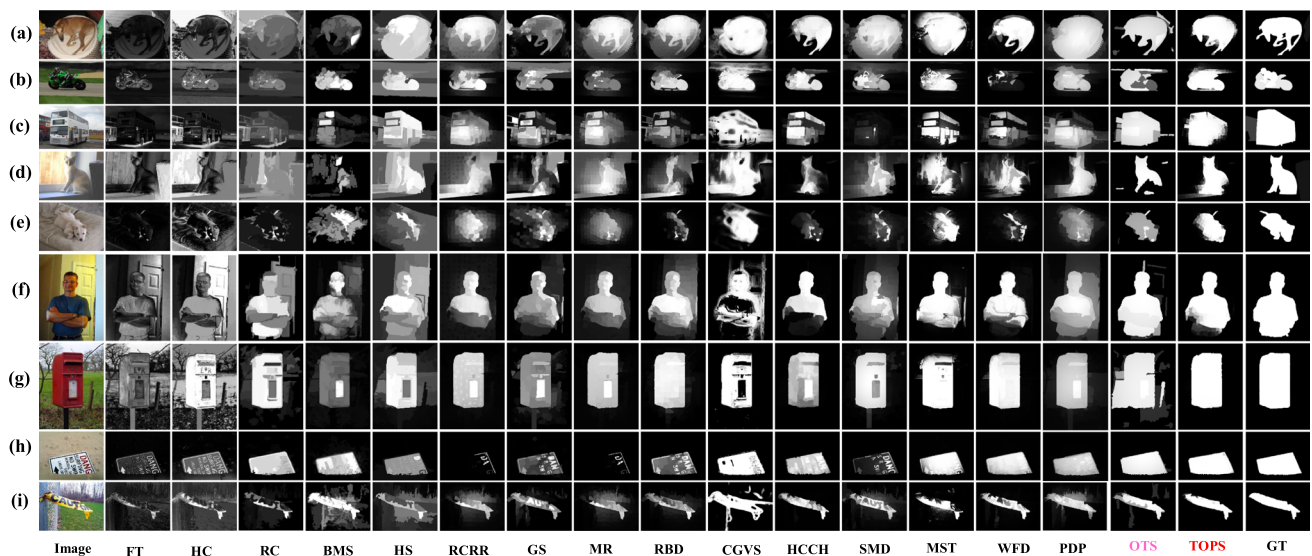
In addition, we introduced another recently proposed metric: the structure-measure,  $Sm.$ , to measure the structure similarity between saliency maps and ground-truth (Fan et al. 2017).

## 4.4 Comparisons with State-of-the-Art Methods on SOD

To qualitatively validate the performance of our methods, our topo-prior OTS and TOPS were compared with current SOTA methods. Visual comparisons are shown in Fig. 8, which illustrates that our TOPS achieves more precise results than the others in some complex and structural scenes. For example, image (a) contains a cluttered background, images (b) and (c) contain complex objects, images (d) and (e) contain objects with low contrast, and the last four images are relatively simple; our TOPS obtains the best results on all of these kinds of images.

The quantitative evaluations are summarized in Fig. 9 and Table 2. The top two rows in Fig. 9 show that our TOPS obtains the best precision–recall results on almost all of the datasets. The F-score curves in the bottom two rows show that our TOPS obtains very high and stable results on most datasets. Table 2 compares another three important metrics, i.e., the MAE, weighted F-measure, and structure-measure. Our TOPS also obtains the best performance on almost all datasets for all of the metrics with a large margin. Further, our method with a single scale (TOPS\_sc) obtains the second best performance (Table 2).

It is worth mentioning that when we do not combine the regional and local features, our topo-saliency prior map (OTS) can still achieve acceptable performance for all of the metrics on all of the datasets; this is indicated by the dashed red lines in Fig. 9 and the values in Table 2. The reason why the proposed OTS is effective is that our topo-complexity



**Fig. 8** Visual comparison of the existing salient object detection methods and our methods (OTS and TOPS) in various scenarios and on different datasets

map captures the spatial and structural prior. The proposed TOPS inherits this advantage from OTS, and at the same time combines the local cues under the guidance of OTS to overcome the limitation of the proposed framework. We also obtain similar performance on DUT-OMRON (Yang et al. 2013) and DUTS-test (Wang et al. 2017b) datasets.

Figure 10 shows that our TOPS obtains similar PR curves to DRFI (Wang et al. 2017a) and better results than HDCT (Kim et al. 2015); meanwhile, TOPS obtains far better MAEs and weighted F-scores on almost all of the datasets (Table 2). However, all of our TOPS and DRFI, HDCT are far worse than the deep learning-based methods. It is worth mentioning that DRFI and HDCT are two top non-DLB supervised methods, and usually serve as comparison baselines for DLB methods. The fully supervised DLB methods are much better than the weakly supervised WSS (Wang et al. 2017b), the non-DLB methods, and the unsupervised methods.

Unlike the DLB methods, our method is unsupervised, which means that we do not need a training process nor expend effort on labeling. To summarize, according to the analyses above, our TOPS achieves SOTA performance among the unsupervised methods.

#### 4.5 Fixation Prediction via Topo-Prior

To quantitatively evaluate the performance of our topological prior on fixation prediction, we adopted four widely used metrics: AUC\_Judd (Judd et al. 2009), AUC\_Borji (Borji et al. 2012), shuffled AUC (sAUC) (Zhang et al. 2008), and normalized scanpath saliency (NSS) (Peters et al. 2005). Experiments were conducted on three popular fixation datasets: ImgSal (Li et al. 2013), MIT (Judd et al. 2009),

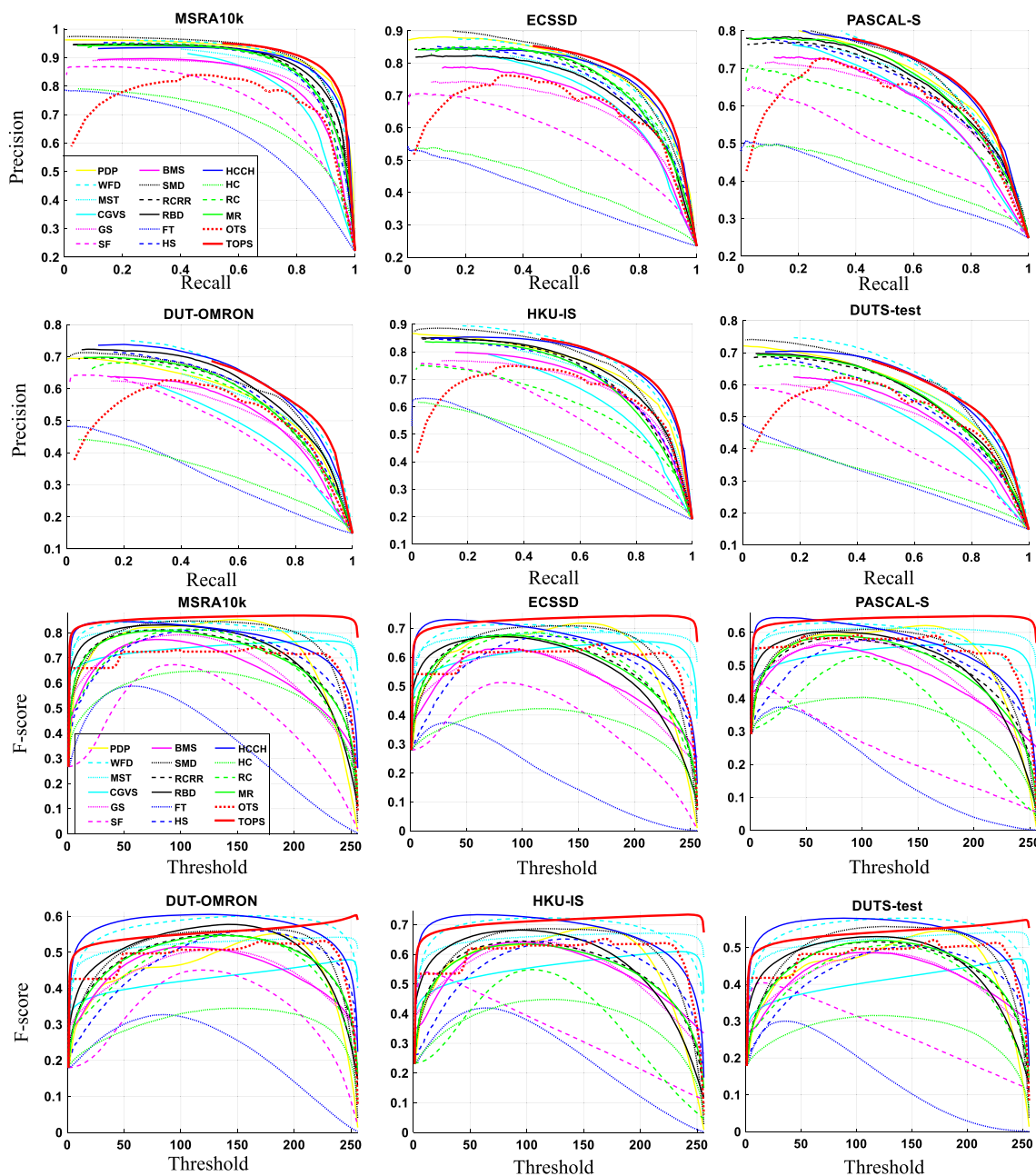
and Toronto (Bruce and Tsotsos 2009). We compared our method with some classical SOTA methods (Itti (Itti et al. 1998), GBVS (Harel et al. 2007), AIM (Bruce and Tsotsos 2009), AWS (Garcia-Diaz et al. 2012), SIG (Hou et al. 2011), and BMS (Zhang and Sclaroff 2015)), some LB methods (Judd (Judd et al. 2009) and eDN (Vig et al. 2014)), and several DLB methods (MLNet (Cornia et al. 2016), Mr-CNN (Liu et al. 2015), and DVA (Wang and Shen 2018)).

In this experiment, the results of GBVS, BMS, Itti, AIM, AWS, SIG, and Judd are adopted from the website<sup>2</sup> of BMS (Zhang and Sclaroff 2015); other results are from their authors' websites or are calculated by their released code with default configurations. The evaluation codes are from the MIT saliency website<sup>3</sup>.

One of the most widely used metrics for fixation prediction is the area under the ROC curve (AUC), which use the human fixation map as ground truth to calculate the *false positive rate* and *true positive rate*. The AUC score is calculated as the area under the ROC drawn by the false positive and true positive rate. AUC\_Judd (Judd et al. 2009) is the classical version of AUC, while AUC\_Borji (Borji et al. 2012) and sAUC (Zhang et al. 2008) are extensions to tackle the influence of center bias. In contrast, NSS (Peters et al. 2005) measures the correspondence between the saliency map and the scanpath. The ImgSal (Li et al. 2013) dataset consists of 235 images collected from 21 observers, with 6 different categories of object size. MIT1003 (Judd et al. 2009) is a large-scale fixation dataset with images collected from 15 observers; it contains 1003 pictures of natural indoor and out-

<sup>2</sup> <http://www.cs.bu.edu/groups/ivc/software/BMS/>

<sup>3</sup> <https://saliency.tuebingen.ai/>



**Fig. 9** Comparison of PR curves and F-measure curves on the MSRA10k (Cheng et al. 2014a), DUT-OMRON (Yang et al. 2013), ECSSD (Shi et al. 2016), PASCAL-S (Yin et al. 2014), HKU-IS (Li and Yu 2015), and DUTS-Test (Wang et al. 2017b) datasets for the task of salient object detection (SOD)

**Table 2** Quantitative comparisons on six salient object detection datasets in terms of MAE, weighted F-measure, and structure-measure. (From the TOPS to the method FT are unsupervised methods; From

DRFI to HDCT are the supervised methods with handcrafted features, while from RASv2 to RFCN are the DLB methods) (Color table online)

Datasets	MSRA10k			HKU-IS			ECSSD			PASCAL-S		
	MAE ↓ <sup>a</sup>	$F_{\beta}^*$ ↑	$S_m$ ↑	MAE	$F_{\beta}^*$	$S_m$	MAE	$F_{\beta}^*$	$S_m$	MAE	$F_{\beta}^*$	$S_m$
<b>TOPS</b>	<b>0.0666<sup>b</sup></b>	<b>0.8191</b>	<b>0.8665</b>	<b>0.1115</b>	<b>0.6606</b>	<b>0.7628</b>	<b>0.1274</b>	<b>0.6727</b>	<b>0.7628</b>	<b>0.1742</b>	<b>0.5759</b>	<b>0.6853</b>
<b>TOPS_sc<sup>c</sup></b>	<b>0.0678</b>	<b>0.8177</b>	<b>0.8611</b>	<b>0.1129</b>	<b>0.6579</b>	<b>0.7585</b>	<b>0.1289</b>	<b>0.6703</b>	<b>0.7526</b>	<b>0.1774</b>	<b>0.5706</b>	<b>0.6773</b>
<b>OTS</b>	0.1280	0.6517	0.7733	0.1532	0.5434	0.7076	0.1846	0.5325	0.6882	0.2022	0.4931	0.6351

**Table 2** continued

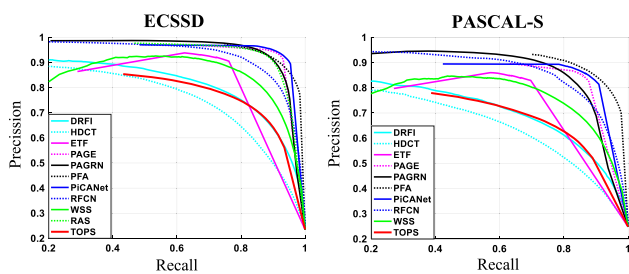
Datasets	MSRA10k		HKU-IS			ECSSD			PASCAL-S			
	(12) <sup>d</sup>	(10)	(12)	(7)	(6)	(8)	(10)	(9)	(11)	(6)	(5)	(6)
HCCH (Liu et al. (2017))	0.0968	0.7287	0.8099	<b>0.1162</b>	<b>0.6028</b>	0.7339	<b>0.1502</b>	0.5840	0.7087	<b>0.1864</b>	0.4888	0.5982
WFD (Huang and Zhang (2018))	<b>0.0906</b>	0.7056	0.8292	0.1221	0.6019	<b>0.7467</b>	0.1513	0.5933	0.7280	0.1895	0.4990	0.6252
MST (Tu et al. (2016))	0.0974	<b>0.7385</b>	0.8098	0.1390	0.5865	0.7145	0.1554	<b>0.6052</b>	0.7122	0.1946	<b>0.5273</b>	0.6354
PDP (Xiao et al. (2018))	0.1144	0.6807	<b>0.8423</b>	0.1722	0.5022	0.7246	0.1841	0.5352	<b>0.7382</b>	0.2237	0.4694	<b>0.6609</b>
SMD (Peng et al. (2016))	0.1040	0.7036	0.8392	0.1559	0.5110	0.7245	0.1739	0.5443	0.7360	0.2095	0.4575	0.6342
CGVS (Yang et al. (2016))	0.1224	0.6829	0.7680	0.1924	0.4965	0.6463	0.1923	0.5426	0.6804	0.2230	0.4647	0.6136
RCRR (Li et al. (2015a))	0.1222	0.6471	0.7902	0.1711	0.4586	0.6768	0.1840	0.4985	0.6941	0.2292	0.4189	0.5992
RBD (Zhu et al. (2014))	0.1080	0.6854	0.8074	0.1424	0.5153	0.7062	0.1714	0.5128	0.6884	0.2028	0.4534	0.6134
HS (Shi et al. (2016))	0.1486	0.6043	0.7866	0.2150	0.4223	0.6742	0.2275	0.4544	0.6851	0.2642	0.4054	0.6165
BMS (Zhang and Sclaroff (2015))	0.1508	0.5664	0.7461	0.1797	0.4410	0.6708	0.1739	0.4346	0.6289	0.2293	0.3688	0.6485
RC (Cheng et al. (2011))	0.1372	0.6080	0.7893	0.1841	0.2915	0.5976	0.1860	0.5091	0.7054	0.3140	0.3134	0.5605
HC (Cheng et al. (2011))	0.2149	0.4814	0.6687	0.2822	0.3047	0.5433	0.3299	0.3085	0.5027	0.3422	0.2951	0.4665
MR (Yang et al. (2013))	0.1255	0.6424	0.7852	0.1782	0.4503	0.6687	0.1862	0.4961	0.6924	0.2240	0.4211	0.6036
GS (Wei et al. (2012))	0.1385	0.6061	0.7751	0.1681	0.4663	0.6908	0.2058	0.4501	0.6608	0.2247	0.4197	0.6039
FT (Achanta et al. (2009))	0.2349	0.3344	0.5673	0.2487	0.2347	0.5044	0.2698	0.1950	0.4481	0.2877	0.1871	0.3973
DRFI (Wang et al. (2017a))	0.1182	0.6537	0.8390	0.1445	0.5063	0.7400	0.1703	0.5191	0.7320	0.2112	0.4344	0.6534
HDCT (Kim et al. (2015))	0.1442	0.5796	0.7962	0.1645	0.4531	0.7089	0.1976	0.4523	0.6858	0.2303	0.3787	0.6136
RASv2 (Chen et al. (2020))	–	–	–	0.0301	0.8938	0.9146	0.0337	0.9132	0.9248	0.0673	0.8199	0.8549
CPD (Wu et al. (2019))	–	–	–	0.0342	0.8747	0.9055	0.0371	0.8980	0.9181	0.0737	0.7998	0.8444
ETF (Zhou and Gu (2020))	–	–	–	0.0676	0.7366	0.7346	0.0824	0.7689	0.7646	0.1159	0.6735	0.6992
PAGE (Wang et al. (2019e))	–	–	–	0.0313	0.0889	0.9034	0.0371	0.9045	0.9120	0.0758	0.8035	0.8378
PFA (Zhao and Wu (2019))	–	–	–	0.0327	0.8754	0.9135	0.0449	0.8708	0.9045	0.0648	0.8035	0.8629
RAS (Chen et al. (2018))	–	–	–	–	–	–	0.0564	0.8569	0.8928	0.1059	0.7350	0.7926
PAGRn (Zhang et al. (2018))	–	–	–	0.0475	0.8195	0.8873	0.0610	0.8335	0.8883	0.0950	0.7339	0.8141
PiCANet (Liu et al. (2018))	–	–	–	0.0308	0.8903	0.9049	0.0345	0.9082	0.9129	0.0667	0.8175	0.8483
WSS (Wang et al. (2017b))	–	–	–	0.0796	0.7079	0.8183	0.1039	0.7092	0.8087	0.1420	0.6092	0.7373
RFCN (Wang et al. (2016))	–	–	–	0.0889	0.6803	0.8557	0.1070	0.6984	0.8518	0.1337	0.6339	0.7980

<sup>a</sup>. The  $\uparrow$  and  $\downarrow$  denote “higher is better” and “lower is better,” respectively.

<sup>b</sup>. Red, green, and blue indicate the top three performances among the *unsupervised* methods, respectively.

<sup>c</sup>. Our method TOPS with a single scale is denoted as *TOPS\_sc*.

<sup>d</sup>. The rank of our method OTS (topo-prior) among the *unsupervised* methods



**Fig. 10** Comparisons with the supervised methods. Our TOPS obtains competitive performance compared with the supervised methods HDCT (Kim et al. 2015) and DRFI (Wang et al. 2017a), but is worse than all of the DLB methods

door scenes. Toronto (Bruce and Tsotsos 2009) contains 120 images and does not contain particular regions of interest.

The comparisons for fixation prediction are exhibited in Table 3. We can see that, although the results of our  $S_{tc}$  (i.e., topo-complexity) are unsatisfactory, the versions of  $S_{tc}$  smoothed by Gaussian filters (i.e.,  $S_{tc\_20}$  and  $S_{tc\_50}$ ) achieve better performance on these datasets, especially  $S_{tc\_20}$ . The reason why the blurred maps perform better than the originals is that although  $S_{tc}$  can indicate the saliency of images, it is hierarchical and sharply regional, which dramatically increases the false positive rate when calculating the AUC; the blurred maps suffer less from this.

After we combined this topological saliency prior with other methods using Eq.5, significant improvements were obtained. Using a blurred maps with  $std = 50$  for promotion (rather than using the  $std = 20$  in Eq.5) enables us to enhance the performance while preserving the advan-

**Table 3** Quantitative comparisons on three fixation prediction datasets in terms of AUC\_Judd, AUC\_Borji, sAUC, and NSS. (From Stc to method SIG are unsupervised method; the Judd and eDN are the conventional supervised methods; and the last three methods are fully supervised DLB methods)

Dataset	ImgSal				MIT1003				Toronto			
	AUC_J.↑ <sup>a</sup>	AUC_B.↑	sAUC↑	NSS↑	AUC_J.	AUC_B.	sAUC	NSS	AUC_J.	AUC_B.	sAUC	NSS
<b>Methods</b>												
Stc	0.7202	0.7158	0.6770	1.1909	0.7621	0.7423	0.6558	1.0995	0.7447	0.7240	0.6101	1.0895
Stc_20	0.8061	0.7809	0.7326	1.4491	0.8041	0.7896	0.6878	1.2684	0.7965	0.7773	0.6334	1.2507
Stc_50	0.8040	0.7889	0.7349	1.3729	0.8056	0.7936	0.6891	1.2708	0.7978	0.7828	0.6258	1.2259
GBVS (Harel et al. (2007))	0.8346	<u>0.8182</u>	0.7527	1.5645	0.8233	0.8134	0.6917	1.3658	0.8316	<u>0.8188</u>	0.6396	1.5194
<b>GBVS*<sup>d</sup></b>	<b><u>0.8388</u></b> <sup>b</sup>	0.8125	<b><u>0.7573</u></b>	<b><u>1.6810</u></b>	<b><u>0.8363</u></b>	<b><u>0.8223</u></b>	<b><u>0.7090</u></b>	<b><u>1.5085</u></b>	<b><u>0.8348</u></b>	0.8156	<b><u>0.6455</u></b>	<b><u>1.5907</u></b>
BMS (Zhang and Sclaroff (2015))	0.7984	0.7783	0.7428	1.5449	0.7846	0.7677	0.7073	1.2287	0.7995	0.7807	<u>0.7095</u>	1.5191
<b>BMS*</b>	<b>0.8245</b> <sup>c</sup>	<b>0.8011</b>	<b>0.7564</b>	<b>1.6621</b>	<b>0.8265</b>	<b>0.8134</b>	<b>0.7197</b>	<b>1.4640</b>	<b>0.8280</b>	<b>0.8078</b>	0.6792	<b>1.5682</b>
Itti (Itti et al. (1998))	0.8018	0.7879	0.7302	1.3770	0.7694	0.7611	0.6769	1.0928	0.8015	0.7907	0.6566	1.2970
<b>Itti*</b>	<b>0.8253</b>	<b>0.8026</b>	<b>0.7499</b>	<b>1.5746</b>	<b>0.8185</b>	<b>0.8067</b>	<b>0.7064</b>	<b>1.3732</b>	<b>0.8241</b>	<b>0.8078</b>	0.6565	<b>1.4627</b>
AIM (Bruce and Tsotsos (2009))	0.7882	0.7713	0.7063	0.8296	0.7792	0.7662	0.6794	0.8125	0.7566	0.7423	0.6689	0.8359
AWS (Garcia-Diaz et al. (2012))	0.7676	0.7548	0.7312	1.3981	0.7533	0.7429	0.7046	1.1250	0.7600	0.7481	0.7043	1.2087
SIG (Hou et al. (2011))	0.7795	0.7648	0.7188	1.3576	0.7590	0.7463	0.6842	1.0859	0.7938	0.7765	0.6959	1.3787
Judd (Judd et al. (2009))	0.7729	0.7613	0.7214	1.2587	0.7525	0.7442	0.6817	1.0056	0.7767	0.7635	0.6745	1.1507
<b>Judd*</b>	<b>0.8157</b>	<b>0.7983</b>	<b>0.7502</b>	<b>1.5586</b>	<b>0.8165</b>	<b>0.8065</b>	<b>0.7105</b>	<b>1.3762</b>	<b>0.8176</b>	<b>0.8029</b>	0.6668	<b>1.4129</b>
eDN (Vig et al. (2014))	–	–	–	–	0.8525	0.8453	0.7181	1.2880	0.8451	0.8353	0.6279	1.2475
<b>eDN*</b>	–	–	–	–	0.8508	<b>0.8457</b>	<b>0.7242</b>	<b>1.5170</b>	0.8398	0.8313	<b>0.6359</b>	<b>1.4723</b>
Mr-CNN (Liu et al. (2015))	–	–	–	–	0.7911	0.7767	0.7294	1.3360	0.7985	0.7761	0.7088	1.4155
<b>Mr-CNN*</b>	–	–	–	–	<b>0.8365</b>	<b>0.8227</b>	<b>0.7463</b>	<b>1.6219</b>	<b>0.8287</b>	<b>0.8051</b>	0.6962	<b>1.5826</b>
MLNet (Cornia et al. (2016))	0.8261	0.7298	0.7034	1.8406	0.8533	0.7723	0.7343	2.2167	0.8492	0.7576	0.6868	1.9967
<b>MLNet*</b>	<b>0.8339</b>	<b>0.7811</b>	<b>0.7425</b>	<b>1.8427</b>	<b>0.8601</b>	<b>0.8272</b>	<b>0.7541</b>	2.0814	0.8434	<b>0.7995</b>	0.6802	1.8335
DVA (Wang and Shen (2018))	–	–	–	–	0.8702	0.8066	0.7561	2.3145	0.8621	0.7806	0.6943	2.1237
<b>DVA*</b>	–	–	–	–	0.8658	<b>0.8358</b>	<b>0.7604</b>	2.1723	0.8502	<b>0.8071</b>	0.6853	1.9707

<sup>a</sup>. “AUC\_J.” and “AUC\_B.” mean “AUC\_Judd” and “AUC\_Borji,” respectively; ↑ means “higher is better.”

<sup>b</sup>. Values in **bold** mean the extended methods are better than the original methods.

<sup>c</sup>. The best performance among the **unsupervised** methods (excluding our topo-complexity method’s results) is denoted by an underline.

<sup>d</sup>. The methods marked with asterisks \* are the extended methods incorporated with our topo-prior

tages of the original methods. The results in Table 3 show that the extended methods, including the unsupervised methods (GBVS\*, BMS\*, and Itti\*) and supervised methods (Judd\*, eDN\*, Mr-CNN\*, MLNet\*, and DVA\*), achieve better results than the original ones for most metrics.

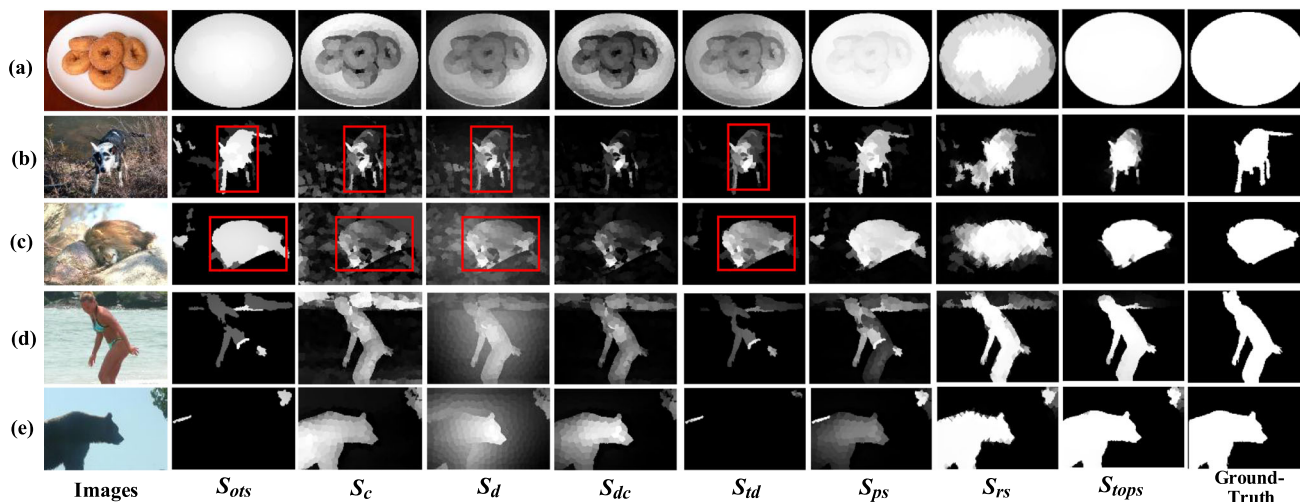
To summarize, although the proposed topo-complexity in Section 3.2.1 is not suitable for predicting fixation directly due to the hierarchical effect, the extended methods adopting Gaussian blurred topo-complexity maps with a proper  $\sigma$  obtained using our strategy have improved performances. This fact further validates the *observation* demonstrated in Section 3.2.2, and reveals the implicitly close relation between the topological features and the visual saliency.

## 4.6 System Mechanism

The mechanism of our system can be visualized in Fig. 11. As the figure shows, different components in the framework contribute differently. Similar to most early unsupervised

methods, our TOPS system also adopts the assumption that the regions near the image borders tend to be the background in topo-prior computation and regional-level cue extraction. Therefore, although our topo-prior ( $S_{or,s}$ ) performs well on images (a), (b), and (c) with cluttered backgrounds, it fails on the images with the objects touching borders, such as images (d) and (e).

However, with the help of the pixel-level processing, the color contrast map ( $S_c$ ) and spatial-color contrast map ( $S_d$ ) are less disturbed by this assumption (see images (d) and (e) in Fig. 11). Along the regional-level processing flow, the regional saliency map ( $S_{r,s}$ ) is obtained by exploiting the relation between the regions in an image and the boundary regions, which has the ability to reduce the image border effect. Therefore, combining these results through Eq. 16 can reduce this kind of risk. The results presented in Figs. 11(d) and (e) illustrate this mechanism. It should be noted that although the strategy is adopted to reduce the image bor-



**Fig. 11** System mechanism. (a) Example with a big target; (b) and (c) images with cluttered backgrounds; (d) and (e) images with objects touching the image borders. The results show that our saliency detec-

tion system can deal with these situations appropriately. The patches in the red rectangles indicate the effectiveness of the weighting of Eq. 10, which make the saliency values of objects more consistent and enhanced

tion effect, such an effect cannot be totally eliminated in our framework. This is one direction of our future work.

Red boxes in Fig. 11 show the effectiveness of the weighting items used in  $S_d$  as computed by Eq. 9. When the topo-prior ( $S_{ots}$ ) successfully detects the objects, appearance contrast can be optimized in the object areas by these two weights of Eq. 10. When topo-prior fails, the problem (2) of  $S_c$  still remains. However, this problem can be solved to a certain extent in the next steps of our model, as shown in Fig. 11, which exhibits the robustness of our model.

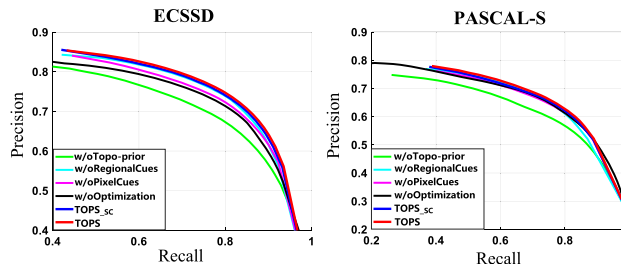
### 4.7 Ablation Analysis

#### 4.7.1 Effectiveness of Various Components

We conducted experiments to evaluate the effectiveness of each component in our framework. In this experiment, we removed each component from the framework successively while keeping the others, and then we compared the final results. To prove the effectiveness of the multi-scale strategy, we also compared the results of the single-scale framework. Figure 12 shows that without topo-prior, the performance degrades dramatically. This indicates that our topo-prior plays a principal role in our method. As shown in Fig. 12, our TOPS performs better than all of the other situations, which demonstrates the effectiveness of our framework.

#### 4.7.2 Replacing Topo-prior with Other Saliency Priors

To validate the proposed topological saliency prior and the saliency detection framework, we compared the performance of our TOPS after replacing the topo-complexity guidance

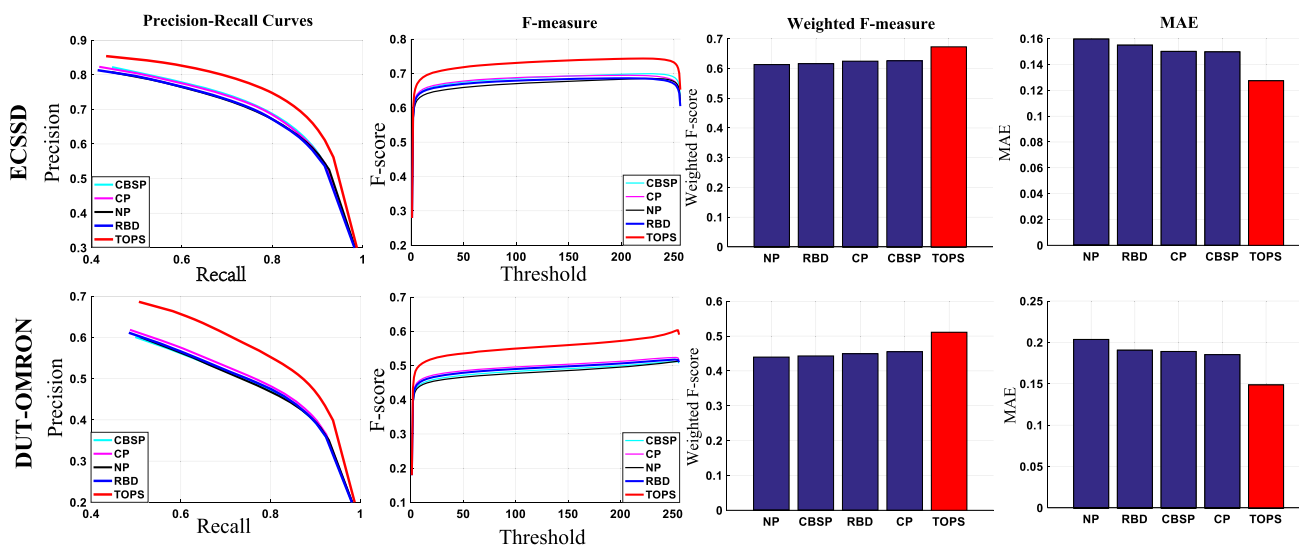


**Fig. 12** Components contribution analysis. The framework without topo-prior, regional-level cues, pixel-level cues, and optimization are denoted as w/oTopo-prior, w/oRegionalCues, w/oPixelCues, and w/oOptimization, respectively. The TOPS with a single scale is denoted as TOPS\_sc

with other saliency priors in our framework while keeping its parameters unchanged in all of the situations.

In this experiment, we replaced our topo-prior with No Prior (NP), the widely used Center Prior (CP), the Contour-Based Spatial Prior (CBSP) (Yang et al. 2016), and the Robust Background Detection prior (RBD) (Zhu et al. 2014). We used the  $S_{bias}$  from Eq.6 to serve as the center prior so as to justify that the performance of TOPS (OTS + selective pathway) on various datasets is mainly determined by the topo-complexity rather than center bias. The CBSP and RBD are two useful saliency priors that can provide reliable spatial and structural prior information for saliency detection, respectively. Similarly, our topological prior, the OTS map, is computed from the UCM (Arbelaez 2006). Consequently, the OTS map inherits precise spatial and structural information from the UCM. We also removed the prior information as a baseline for other situations, which is denoted as no prior (NP). Figure 13 compares the results of various situa-





**Fig. 13** Comparisons of our framework guided by different saliency priors on ECSSD and DUT-OMRON

tions guiding the selective pathway with NP, CP, CBSP, or RBD, and the results of our TOPS (i.e., OTS + selective pathway) for four metrics (PR curves, F-measure, weighted-F, and MAE) on two representative challenging datasets DUT-OMRON (Yang et al. 2013) and ECSSD (Shi et al. 2016). From this figure, we can see that our topo-complexity-guided method (TOPS) outperforms all of the others in all of the metrics and on all of the datasets, by a large margin.

Compared with the baseline (the NP + selective pathway), most of the other prior-guided methods obtain better performance, which means that prior information can indeed improve performance on the saliency detection task. The comparison of TOPS and the no prior (NP) solution validates the significant effect of our topological features. More importantly, that our TOPS outperforms CBSP and RBD indicates that our topological saliency prior can obtain more precise spatial and structural information for saliency detection.

Consequently, we conclude that the topo-complexity derived from image topological features is a promising prior that is quite suitable for the task of saliency detection.

#### 4.7.3 Promotion for Existing Unsupervised Methods

Our topo-prior (OTS) not only performs well in the proposed framework, but it can also promote the performance of existing unsupervised methods when replacing their priors with OTS. For example, LMLC (Xie et al. 2012) uses a convex hull of points of interest as the middle level prior to facilitating the inference of Bayesian saliency at each pixel. Meanwhile, CGVS (Yang et al. 2016) adopts the contour information as spatial prior to guide saliency detection. In this experiment, we directly replaced the middle level prior with our OTS for LMLC and a proper Gaussian blurred OTS with an std

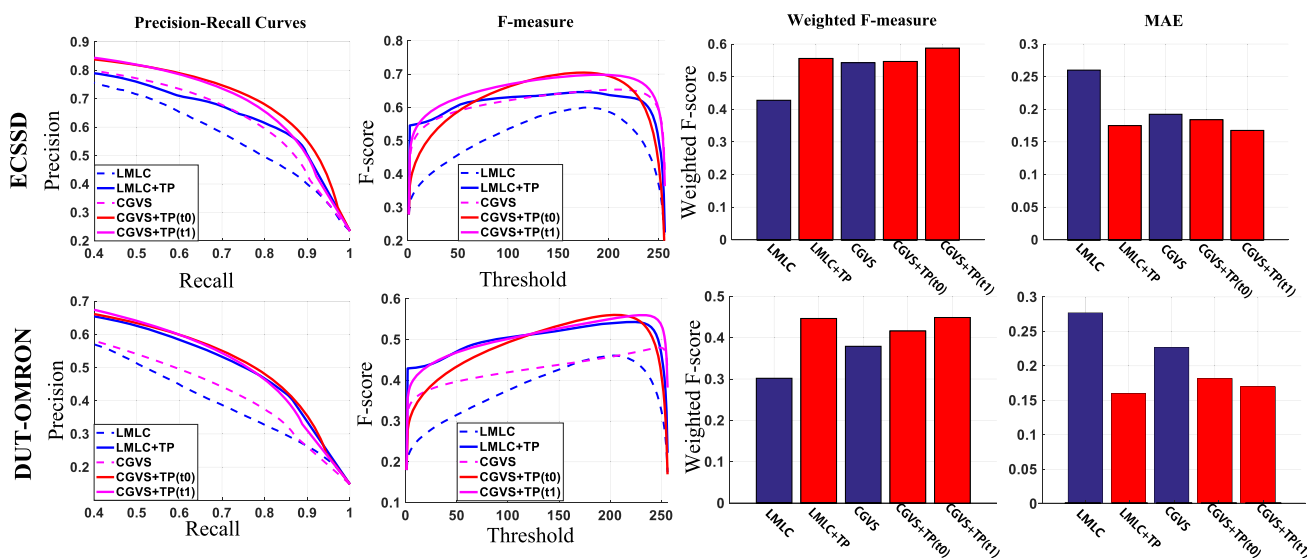
of  $\sigma = 20$  for CGVS. The results shown in Fig. 14 indicate that our topo-prior significantly promotes their performances (other method + TP) in terms of all of the metrics on both ECSSD and DUT-OMRON datasets. These findings further validate the superiority of our topological prior.

#### 4.7.4 Promotion for Supervised and DNN-based Methods

Generally, supervised methods employ annotations and many kinds of handcrafted or deep features, and usually obtain better results than unsupervised methods. However, to the best of our knowledge, very few of them exploit topological properties (Huang et al. 2009; Gu et al. 2013; Chen et al. 2019). Consequently, in this section, we applied our topological saliency prior ( $S_{ots}$ ) to supervised methods DRFI (Wang et al. 2017a) and HDCT (Kim et al. 2015), the unsupervised deep neural network (DNN)-based method HCA (Qin et al. 2018), the supervised DNN-based method ETF (Zhou and Gu 2020), and two newly proposed end-to-end DLB methods CPD (Wu et al. 2019) and RASv2 (Chen et al. 2020).

According to the details of DRFI (Wang et al. 2017a), the features of backgroundness ( $b_{1\sim 29}$ ) play the most important role in DRFI. Therefore, we simply modified these features by  $b_{1\sim 29}^* = b_{1\sim 29} \cdot (1 + S_{ots})$ , and the rest of DRFI remained unchanged. According to the details of HDCT (Kim et al. 2015), a global salient region map obtained by HDCT and a local salient region map obtained via a random forest are combined to compute the final results. We simply multiplied our topo-prior map with both of these salient maps directly, and the rest of HDCT (Kim et al. 2015) remained unchanged.

According to the details of HCA (Qin et al. 2018), the deep features extracted from the pre-trained VGG network are processed by hierarchical cellular automata. Note that



**Fig. 14** Comparisons with other methods (including LMLC (Xie et al. 2012), and CGVS (Yang et al. 2016)) on two datasets when replacing their priors with the proposed topo-prior (OTS).  $t_0$  and  $t_1$  mean the situations of 1 and 2 iterations, respectively

HCA also employs a scheme to integrate with saliency priors. Therefore, we simply adopted our topo-prior as this prior information, and the rest of HCA remained unchanged. According to the details of ETF (Zhou and Gu 2020), a topological feature map  $S_{tf}$  is incorporated into a DNN. We simply modified the topo-feature maps by our topo-prior as  $S_{tf}^* = S_{tf} \cdot (1 + S_{ots})$ , and the rest of ETF remained unchanged.

Integrating existing conventional methods into the SOTA deep neural networks while obtaining better performance is also meaningful and practical. One of the difficulties is that they are end-to-end and too compact to integrate with other handcrafted cues. Two newly proposed state-of-the-arts, e.g., CPD (Wu et al. 2019), and RASv2 (Chen et al. 2020) can be appropriately integrated into by our topo-prior, since both of them generate prior maps (i.e., attention maps,  $S_{att}$ ) to guide the rest of the models to produce better results.

Similar to previous modulated methods (DRFI\*, HDCT\*, HCA\*, and ETF\*), the core idea is to modify the generated prior map ( $S_{att}$ ) with our topo-prior ( $S_{ots}$ ) with a learnable weighting ( $\omega$ ), which is formulated as  $S_{att}^* = S_{att} \cdot (1 + \omega \cdot S_{ots})$ . The rests of CPD and RASv2 remained unchanged. For RASv2, we used the prior map generated by the last layer as  $S_{att}$ . One difference compared with the previous modulated methods is that we *retrained* the modulated models.

From Table 4 we can see that the topo-prior modulated CPD (i.e., CPD\*) performs consistently better across all the three datasets in terms of both metrics than the original CPD. In addition, compared with the original RASv2, the RASv2\* performs better on the PASCAL-S dataset, while slightly worse on other datasets. We argue with two probably reasons for this observation. First, the topological features like ours

may have been partly learned in some SOTA methods such as RASv2, while other methods like CPD does not. Second, PASCAL-S is a special datasets compared with other SOD datasets, because the ground-truths were generated by the fixation and full segmentations rather than directly outlining the salient objects by person. This fact makes the PASCAL-S one of the hardest SOD datasets, and also the one that is most consistent with human cognition. Fortunately, our topo-prior can appropriately capture the fixation as Table 3 shows, which makes our topo-prior able to improve the RASv2 on difficult datasets such as PASCAL-S.

The results (especially the DRFI\*, HDCT\*, HCA\*, ETF\*, CPD\*, and RASv2\*) in Table 4 show that our topo-prior has large potential to promote the performance of the supervised, unsupervised DNN-based methods, and even the newly proposed SOTA DLB methods, which further confirms the value of our topo-prior.

### 4.8 Time Efficiency

Because our system is based on the UCMs and superpixel map, the complexity is affected by the number, area, and structure of the regions. To further demonstrate the efficiency of our method, we show the average time cost for each component of our framework in Table 5 and for each image in Table 6.

Our system is implemented in Matlab2016a using a PC with 3.60 GHz CPU and 32G RAM. As shown in Table 5, it takes about 0.7 s to obtain a UCM, 1.7 s to calculate the topological saliency prior  $S_{ots}$ , 2.4 s to compute the pixel-level saliency  $S_{ps}$ , 0.5 s for the regional-level saliency  $S_{rs}$ , and another 0.3 s for saliency fusion and optimization. It

**Table 4** Quantitative validation of the promotion for supervised SOD methods and DNN-based methods combined with our topo-prior. DRFI and HDCT are the supervised methods with handcrafted features; HCA

and ETF are classical DNN-based methods; CPD and RASv2 are the newly proposed end-to-end DLB SOTA methods

Datasets Methods	ECSSD		PASCAL-S		HKU-IS	
	MAE ↓	$F_{\beta}^*$ ↑	MAE	$F_{\beta}^*$	MAE	$F_{\beta}^*$
DRFI ( Wang et al. (2017a))	0.170	0.519	0.211	0.434	0.145	0.506
<b>DRFI*</b>	<b>0.169</b> <sup>Δ</sup>	<b>0.543</b>	<b>0.208</b>	<b>0.475</b>	0.154	<b>0.511</b>
HDCT ( Kim et al. (2015))	0.198	0.452	0.230	0.379	0.165	0.453
<b>HDCT*</b>	<b>0.191</b>	<b>0.489</b>	<b>0.222</b>	<b>0.421</b>	<b>0.158</b>	<b>0.488</b>
HCA ( Qin et al. (2018))	0.119	0.675	0.160	0.591	0.115	0.627
<b>HCA*</b>	<b>0.116</b>	<b>0.684</b>	<b>0.158</b>	0.582	<b>0.103</b>	<b>0.645</b>
ETF ( Zhou and Gu (2020))	0.082	0.769	0.116	0.674	0.068	0.737
<b>ETF*</b>	<b>0.078</b>	<b>0.796</b>	<b>0.112</b>	<b>0.695</b>	<b>0.061</b>	<b>0.771</b>
CPD ( Wu et al. (2019))	0.037	0.898	0.074	0.800	0.034	0.875
<b>CPD*</b>	<b>0.035</b>	<b>0.907</b>	<b>0.069</b>	<b>0.809</b>	<b>0.031</b>	<b>0.888</b>
RASv2 ( Chen et al. (2020))	0.034	0.913	0.067	0.820	0.030	0.894
<b>RASv2*</b>	0.034	0.909	<b>0.065</b>	<b>0.822</b>	0.031	0.891

<sup>Δ</sup>. Values in **bold** denote that the extended methods with our topo-priors are better than the originals

**Table 5** Time Efficiency Analysis for Each Component

Component	ucm	S <sub>ots</sub>	S <sub>ps</sub>	S <sub>rs</sub>	opt.	TOPS <sub>sc</sub>	TOPS
Time (s)	0.7	1.7	2.4	0.5	0.3	3.2	5.5

\*. The time for computation of **ucm** is included in S<sub>ots</sub>, and the time for computing S<sub>ots</sub> is included in S<sub>ps</sub>

**Table 6** Time efficiency comparison with other methods

Method	PDP	HCCH	DRFI	HDCT	TOPS <sub>sc</sub>	TOPS
Time (s)	2.5	2.9	5.9	4.1	<b>3.2</b>	<b>5.5</b>

\*. The code of all methods listed here are written in MATLAB.

\*. Training times for DRFI and HDCT are not included

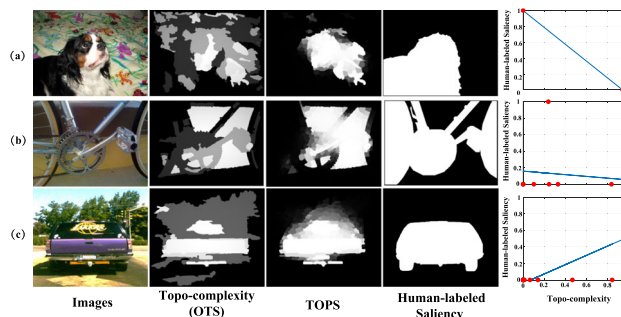
costs about 3.2 s for single-scale computation and around 5.5 s (with three threads) for the whole system.

Since our multi-scale version can use multiple threads for acceleration, we use three threads to calculate each scale. As shown in Table 6, our method with a single scale costs less time than DRFI and HDCT, but our multi-scale version costs more. However, both of our single- and multi-scale methods outperform all of the other unsupervised methods, as shown in Table 2.

### 4.9 Limitations and Future Work

Our system extracts topological information based on a segmentation method and combines the color contrast map under the guidance of topo-prior. Therefore, there exist two limitations affecting the performance.

Firstly, the topo-prior may be not good. Statistics conducted in Sec.3.2.2 have shown that positive relationship exists between topo-complexity and visual saliency. However,



**Fig. 15** Examples of the cases of negative correlation between topo-complexity (i.e., topo-prior, OTS) and human-labeled saliency. (a) and (b) show the failure examples caused by bad topo-priors while (c) is an imperfect example caused by insufficient low-level segmentation scheme

there are still very few images having negative relation. Figure 15 (a) shows an image containing a cluttered background

and (b) shows an image containing an object with large but sparse shape.

Methods only employing the bottom-up cues regard the regions with higher local contrast more salient, then it will assign higher saliency values to the regions of higher contrast, even they are located in background areas. Our topo-complexity (or OTS) is purely bottom-up based and suffered from this problem, as shown in the second column of Fig. 15. We argue that there may exist higher level or top-down modulation while we are paying attention to such scenes, for example, semantic, or task-related modulation.

Secondly, the topo-prior is good but the whole system can not segment the objects accurately. Although the object shown in Fig. 15 (c) can be roughly located, the final segmentation result is unfavorable because of the low contrast; and the low-level segmentation used in our method is insufficient to segment the targets even the topological prior is almost correct.

To overcome these limitations, we plan to incorporate higher-level priors into the non-selective pathway and more sensitive local cues into the selective pathway of our framework.

Moreover, our model is enclosed-region based and needs off-line low-level feature (i.e., contour) extraction, which makes it seems difficult to combine with existing networks. However, after extensive experiments, we found that our assumption of enclosed region does not degrade our model's performance too much, because in real-world scenes, enclosed regions are quite common (Zitnick and Dollár 2014; Cheng et al. 2014b), or at least, for object-based computer vision tasks (e.g., object detection and object tracking), our targets are enclosed-region shaped.

As for the requirement of extracting low-level features off-line, we admit this makes our model not as easy as those end-to-end network models in implementation. However, we think this also leave us chance to employ effective knowledge-based models to extract effective low-level (and even middle- and high-level) features when no large amount of samples are available for those end-to-end network models. In fact, many researches have demonstrated that effective knowledge-based processing (e.g., pose estimation, simi-global matching) can largely improve the performance of end-to-end network models (Klingner et al. 2020; Seki and Pollefeys 2017). Our future work is to build more effective low-level feature extraction (i.e., contour extraction) used for our topological complexity computation.

## 5 Conclusion

In this article, inspired by Topological Perception Theory (TPT), we established a relatively complete and effective system for saliency detection. It contains a computational model

of TPT, a topo-complexity calculation method, and a topo-prior-guided framework inspired by Guided Search Theory for combining global–local saliency. The computational TPT model provides a compact mathematical solution for topological feature extraction. The topo-complexity calculation method offers a feasible scheme to compute the topological saliency prior maps from real-world images. The framework combining local cues under the guidance of the topological saliency prior obtains better performance than unsupervised state-of-the-art methods. The topo-complexity prior can be used to predict human fixation and promote other methods. Moreover, our extensive analyses confirm that the topo-prior is effective and quite suitable to act as a novel saliency prior that can also be incorporated into existing salient object detection models for performance promotion. Qualitative and quantitative comparisons confirm that our system can achieve competitive performance in SOD tasks compared with even the newly proposed SOTA methods. Besides, the close relation revealed in this work between the topological properties and the visual saliency provides a significant perspective for further exploiting the connection of saliency detection and topological perception theory.

**Acknowledgements** The authors would like to thank Professor Lin Chen for his helpful discussions and suggestions on the modeling of his topological perception theory. This work was supported by the Key Area R&D Program of Guangdong Province (#2018B030338001), the Natural Science Foundations of China (#62076055, #61806041). This work was also supported by the 111 Project (B12027) of China. We also thank LetPub for its linguistic assistance during the preparation of this manuscript.

## References

- Achanta R, Hemami S, Estrada F, Susstrunk S (2009) Frequency tuned salient region detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 1597–1604
- Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., & Süsstrunk, S. (2012). Slic superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(11), 2274–2282.
- Adelson, E. H., Anderson, C. H., Bergen, J. R., Burt, P. J., & Ogden, J. M. (1984). Pyramid methods in image processing. *RCA engineer*, 29(6), 33–41.
- Arbelaez P (2006) Boundary extraction in natural images using ultrametric contour maps. In: *Conference on IEEE Conference on Computer Vision and Pattern Recognition Workshop*, pp 182–182
- Borji A (2019) Saliency prediction in the deep learning era: Successes and limitations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*
- Borji A, Itti L (2011) Scene classification with a sparse set of salient regions. *IEEE International Conference on Robotics and Automation* pp 1902–1908
- Borji, A., & Itti, L. (2012). State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1), 185–207.

- Borji, A., Sihite, D. N., & Itti, L. (2012). Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, 22(1), 55–69.
- Borji A, Cheng MM, Hou Q, Jiang H, Li J (2014) Salient object detection: A survey. *Computational Visual Media* pp 1–34
- Bruce, N. D., & Tsotsos, J. K. (2009). Saliency, attention, and visual search: An information theoretic approach. *Journal of vision*, 9(3), 5.
- Chen J, Li Q, Wu W, Ling H, Wu L, Zhang B, Li P (2019) Saliency detection via topological feature modulated deep learning. *2019 IEEE International Conference on Image Processing* pp 1630–1634
- Chen, L. (1982). Topological structure in visual perception. *Science*, 218(4573), 699–700.
- Chen, L. (2005). The topological approach to perceptual organization. *Visual Cognition*, 12(4), 553–637.
- Chen, L., Zhang, S., & Mandyam, V. S. (2003). Global perception in small brains: topological pattern recognition in honey bees. *Proceedings of the National Academy of Sciences of the United States of America*, 100(11), 6884–6889.
- Chen, S., Zheng, L., Hu, X., & Zhou, P. (2016). Discriminative saliency propagation with sink points. *Pattern recognition*, 60, 2–12.
- Chen S, Tan X, Wang B, Hu X (2018) Reverse attention for salient object detection. In: *Proceedings of the European Conference on Computer Vision (ECCV)*, pp 234–250
- Chen, S., Tan, X., Wang, B., Lu, H., Hu, X., & Fu, Y. (2020). Reverse attention-based residual network for salient object detection. *IEEE Transactions on Image Processing*, 29, 3763–3776.
- Chen X, Zheng A, Li J, Lu F (2017) Look, perceive and segment: Finding the salient objects in images via two-stream fixation-semantic cnns. In: *Proceedings of the IEEE International Conference on Computer Vision*, pp 1050–1058
- Cheng MM, Zhang GX, Mitra NJ, Huang X, Hu SM (2011) Global contrast based salient region detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 409–416
- Cheng, M. M., Mitra, N. J., Huang, X., Torr, P. H., & Hu, S. M. (2014a). Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3), 569–582.
- Cheng MM, Zhang Z, Lin WY, Torr P (2014b) Bing: Binarized normed gradients for objectness estimation at 300fps. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 3286–3293
- Cong, R., Lei, J., Fu, H., Cheng, M. M., Lin, W., & Huang, Q. (2019). Review of visual saliency detection with comprehensive information. *IEEE Transactions on Circuits and Systems for Video Technology*, 29(10), 2941–2959.
- Cornia M, Baraldi L, Serra G, Cucchiara R (2016) A deep multi-level network for saliency prediction. In: *2016 23rd International Conference on Pattern Recognition (ICPR)*, pp 3488–3493
- DeYoe, E. A., & Van Essen, D. C. (1988). Concurrent processing streams in monkey visual cortex. *Trends in neurosciences*, 11(5), 219–226.
- Fan DP, Cheng MM, Liu Y, Li T, Borji A (2017) Structure-measure: A new way to evaluate foreground maps. In: *IEEE International Conference on Computer Vision*, pp 4548–4557
- Fang, Y., Wang, J., Narwaria, M., Le Callet, P., & Lin, W. (2014). Saliency detection for stereoscopic images. *IEEE Transactions on Image Processing*, 23(6), 2625–2636.
- Gao, Y., Shi, M., Tao, D., & Xu, C. (2015). Database saliency for fast image retrieval. *IEEE Transactions on Multimedia*, 17(3), 359–369.
- Garcia-Diaz, A., Fdez-Vidal, X. R., Pardo, X. M., & Dosil, R. (2012). Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1), 51–64.
- Gong C, Tao D, Liu W, Maybank SJ, Fang M, Fu K, Yang J (2015) Saliency propagation from simple to difficult. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 2531–2539
- Gu, X., Fang, Y., & Wang, Y. (2013). Attention selection using global topological properties based on pulse coupled neural network. *Computer Vision Image Understanding*, 117(10), 1400–1411.
- Harel J, Koch C, Perona P (2007) Graph-based visual saliency. In: *Advances in Neural Information Processing Systems*, pp 545–552
- He, L., Zhou, K., Zhou, T., He, S., & Chen, L. (2015). Topology-defined units in numerosity perception. *Proceedings of the National Academy of Sciences*, 112(41), E5647–E5655.
- He S, Tavakoli HR, Borji A, Mi Y, Pugeault N (2019) Understanding and visualizing deep visual saliency models. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 10206–10215
- Heeger DJ, Bergen JR (1995) Pyramid-based texture analysis/synthesis. In: *the 22nd annual conference on Computer Graphics and Interactive Techniques*, Citeseer, pp 229–238
- Hornung A, Pritch Y, Krahenbuhl P, Perazzi F (2012) Saliency filters: Contrast based filtering for salient region detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 733–740
- Hou X, Zhang L (2007) Saliency detection: A spectral residual approach. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 1–8
- Hou, X., Harel, J., & Koch, C. (2011). Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1), 194–201.
- Huang, X., & Zhang, Y. (2018). Water flow driven salient object detection at 180 fps. *Pattern Recognition*, 76, 95–107.
- Huang, X., & Zhang, Y. J. (2017). 300-fps salient object detection via minimum directional contrast. *IEEE Transactions on Image Processing*, 26(9), 4243–4254.
- Huang Y, Huang K, Tan T, Tao D (2009) A novel visual organization based on topological perception. In: *Asian Conference on Computer Vision*, Springer, pp 180–189
- Itti, L., & Koch, C. (2001). Computational modelling of visual attention. *Nature Reviews Neuroscience*, 2(3), 194–203.
- Itti, L., Koch, C., & Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11, 1254–1259.
- Ji, Q., Fang, Z., Xie, Z., & Lu, Z. (2013). Video abstraction based on the visual attention model and online clustering. *Signal Processing-Image Communication*, 28(3), 241–253.
- Jiang H, Wang J, Yuan Z, Wu Y, Zheng N, Li S (2013a) Salient object detection: A discriminative regional feature integration approach. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 2083–2090
- Jiang P, Ling H, Yu J, Peng J (2013b) Salient region detection by ufo: Uniqueness, focusness and objectness. In: *IEEE International Conference on Computer Vision*, pp 1976–1983
- Judd T, Ehinger K, Durand F, Torralba A (2009) Learning to predict where humans look. In: *IEEE International Conference on Computer Vision*, IEEE, pp 2106–2113
- Kim, J., Han, D., Tai, Y. W., & Kim, J. (2015). Salient region detection via high-dimensional color transform and local spatial support. *IEEE transactions on image processing*, 25(1), 9–23.
- Klingner M, Termöhlen JA, Mikolajczyk J, Fingscheidt T (2020) Self-supervised monocular depth estimation: Solving the dynamic object problem by semantic guidance. In: *European Conference on Computer Vision*, Springer, pp 582–600
- Koch C, Ullman S (1987) Shifts in selective visual attention: towards the underlying neural circuitry. In: *Matters of intelligence*, Springer, pp 115–141
- Koffka K Principles of Gestalt psychology *Principles of Gestalt psychology*. Routledge

- Kruthiventi, S. S., Ayush, K., & Babu, R. V. (2017). Deepfix: A fully convolutional neural network for predicting human eye fixations. *IEEE Transactions on Image Processing*, 26(9), 4446–4456.
- Kummerer M, Wallis TSA, Gatys LA, Bethge M (2017) Understanding low- and high-level contributions to fixation prediction. In: *The IEEE International Conference on Computer Vision (ICCV)*
- Li C, Yuan Y, Cai W, Xia Y, Dagan Feng D (2015a) Robust saliency detection via regularized random walks ranking. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 2710–2717
- Li G, Yu Y (2015) Visual saliency based on multiscale deep features. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 5455–5463
- Li G, Xie Y, Wei T, Wang K, Lin L (2018) Flow guided recurrent neural encoder for video salient object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 3243–3252
- Li, J., Levine, M. D., An, X., Xu, X., & He, H. (2013). Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4), 996–1010.
- Li N, Sun B, Yu J (2015b) A weighted sparse coding framework for saliency detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 5216–5223
- Lin, X., Wang, Z. J., Ma, L., & Wu, X. (2019). Saliency detection via multi-scale global cues. *IEEE Transactions on Multimedia*, 21(7), 1646–1659.
- Liu N, Han J, Zhang D, Wen S, Liu T (2015) Predicting eye fixations using convolutional neural networks. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 362–370
- Liu N, Han J, Yang MH (2018) Picanet: Learning pixel-wise contextual attention for saliency detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 3089–3098
- Liu, Q., Hong, X., Zou, B., Chen, J., Chen, Z., & Zhao, G. (2017). Hierarchical contour closure-based holistic salient object detection. *IEEE Transactions on Image Processing*, 26(9), 4537–4552.
- Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., et al. (2010). Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(2), 353–367.
- Livingstone, M. S., & Hubel, D. H. (1987). Psychophysical evidence for separate channels for the perception of form, color, movement, and depth. *Journal of Neuroscience*, 7(11), 3416–3468.
- Ma, C., Miao, Z., Zhang, X. P., & Li, M. (2017). A saliency prior context model for real-time object tracking. *IEEE Transactions on Multimedia*, 19(11), 2415–2424.
- Marr, David (1982) Vision: A computational investigation into the human representation and processing of visual information. *Quarterly Review of Biology* 8
- Otsu, N. (1979). A threshold selection method from gray-level histograms. *IEEE Transactions on Systems, Man, and Cybernetics*, 9(1), 62–66.
- Peng H, Li B, Xiong W, Hu W, Ji R (2014) Rgb-d salient object detection: A benchmark and algorithms. In: *European Conference on Computer Vision*, Springer, pp 92–109
- Peng, H., Li, B., Ling, H., Hu, W., Xiong, W., & Maybank, S. J. (2016). Salient object detection via structured matrix decomposition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(4), 818–832.
- Perazzi F, Krähenbühl P, Pritch Y, Hornung A (2012) Saliency filters: Contrast based filtering for salient region detection. In: *2012 IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, pp 733–740
- Peters, R. J., Iyer, A., Itti, L., & Koch, C. (2005). Components of bottom-up gaze allocation in natural images. *Vision research*, 45(18), 2397–2416.
- Qin Y, Lu H, Xu Y, Wang H (2015) Saliency detection via cellular automata. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 110–119
- Qin, Y., Feng, M., Lu, H., & Cottrell, G. W. (2018). Hierarchical cellular automata for visual saliency. *International Journal of Computer Vision*, 126(7), 751–770.
- Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., & Yang, Q. (2017). Rgb-d salient object detection via deep fusion. *IEEE Transactions on Image Processing*, 26(5), 2274–2285.
- Rahtu E, Kannala J, Salo M, Heikkilä J (2010) Segmenting salient objects from images and videos. *European Conference on Computer Vision* pp 366–379
- Scharfenberger C, Wong A, Fergani K, Zelek JS, Clausi DA (2013) Statistical textural distinctiveness for salient region detection in natural images. In: *IEEE Conference on Computer Vision and Pattern Recognition*
- Seki A, Pollefeys M (2017) Sgm-nets: Semi-global matching with neural networks. In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 231–240
- Shi, J., Yan, Q., Xu, L., & Jia, J. (2016). Hierarchical image saliency detection on extended cssd. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(4), 717–729.
- Siva P, Russell C, Xiang T, Agapito L (2013) Looking beyond the image: Unsupervised learning for object saliency and detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 3238–3245
- Song, H., Liu, Z., Du, H., Sun, G., Le Meur, O., & Ren, T. (2017). Depth-aware salient object detection and segmentation via multi-scale discriminative saliency fusion and bootstrap learning. *IEEE Transactions on Image Processing*, 26(9), 4204–4216.
- Treisman, A. M., & Gelade, G. (1980). A feature-integration theory of attention. *Cognitive Psychology*, 12(1), 97–136.
- Tu WC, He S, Yang Q, Chien SY (2016) Real-time salient object detection with a minimum spanning tree. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 2334–2342
- Vig E, Dorr M, Cox D (2014) Large-scale optimization of hierarchical features for saliency prediction in natural images. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 2798–2805
- Wang, B., Zhou, T. G., Zhuo, Y., & Chen, L. (2007). Global topological dominance in the left hemisphere. *Proceedings of the National Academy of Sciences*, 104(52), 21014–21019.
- Wang, J., Jiang, H., Yuan, Z., Cheng, M. M., Hu, X., & Zheng, N. (2017a). Salient object detection: A discriminative regional feature integration approach. *International Journal of Computer Vision*, 123(2), 251–268.
- Wang L, Wang L, Lu H, Zhang P, Ruan X (2016) Saliency detection with recurrent fully convolutional networks. In: *European Conference on Computer Vision*, Springer, pp 825–841
- Wang L, Lu H, Wang Y, Feng M, Wang D, Yin B, Ruan X (2017b) Learning to detect salient objects with image-level supervision. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 136–145
- Wang T, Zhang L, Wang S, Lu H, Yang G, Ruan X, Borji A (2018) Detect globally, refine locally: A novel approach to saliency detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 3127–3135
- Wang, W., & Shen, J. (2018). Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5), 2368–2378.
- Wang W, Lai Q, Fu H, Shen J, Ling H (2019a) Salient object detection in the deep learning era: An in-depth survey. arXiv preprint [arXiv:1904.09146](https://arxiv.org/abs/1904.09146)
- Wang W, Shen J, Cheng MM, Shao L (2019b) An iterative and cooperative top-down and bottom-up inference network for salient object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 5968–5977
- Wang W, Shen J, Dong X, Borji A, Yang R (2019c) Inferring salient objects from human fixations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp 1

- Wang W, Shen J, Xie J, Cheng MM, Ling H, Borji A (2019d) Revisiting video saliency prediction in the deep learning era. *IEEE Transactions on Pattern Analysis and Machine Intelligence* pp 1
- Wang W, Zhao S, Shen J, Hoi SC, Borji A (2019e) Salient object detection with pyramid attention and salient edges. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 1448–1457
- Wei Y, Wen F, Zhu W, Sun J (2012) Geodesic saliency using background priors. In: *European Conference on Computer Vision*, Springer, pp 29–42
- Wolfe JM (1994) Guided search 2.0 a revised model of visual search. *Psychon Bull Rev* 1(2):202–238
- Wolfe, J. M., Melissa, L.-H. V., Evans, K. K., & Greene, M. R. (2011). Visual search in scenes involves selective and nonselective pathways. *Trends in Cognitive Sciences*, 15(2), 77–84.
- Wu Z, Su L, Huang Q (2019) Cascaded partial decoder for fast and accurate salient object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 3907–3916
- Xiao, X., Zhou, Y., & Gong, Y. J. (2018). RGB-D saliency detection with pseudo depth. *IEEE Transactions on Image Processing*, 28(5), 2126–2139.
- Xie, Y., Lu, H., & Yang, M. H. (2012). Bayesian saliency via low and mid level cues. *IEEE Transactions on Image Processing*, 22(5), 1689–1698.
- Yan Q, Xu L, Shi J, Jia J (2013) Hierarchical saliency detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 1155–1162
- Yang C, Zhang L, Lu H, Ruan X, Yang MH (2013) Saliency detection via graph-based manifold ranking. *IEEE Conference on Computer Vision and Pattern Recognition* pp 3166–3173
- Yang KF, Gao X, Zhao JR, Li YJ (2015) Segmentation-based salient object detection. In: *CCF Chinese Conference on Computer Vision*, pp 94–102
- Yang, K. F., Li, H., Li, C. Y., & Li, Y. J. (2016). A unified framework for salient structure detection by contour-guided visual search. *IEEE Transactions on Image Processing*, 25(8), 3475–3488.
- Yin L, Hou X, Koch C, Rehg JM, Yuille AL (2014) The secrets of salient object segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 280–287
- Yuan, Y., Li, C., Kim, J., Cai, W., & DD F. (2017). Reversion correction and regularized random walk ranking for saliency detection. *IEEE Transaction Image Process*, 27(3), 1–1.
- Zeng Y, Zhuge Y, Lu H, Zhang L, Qian M, Yu Y (2019) Multi-source weak supervision for saliency detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 6074–6083
- Zhang J, Sclaroff S (2013) Saliency detection: A boolean map approach. In: *Proceedings of the IEEE international conference on computer vision*, pp 153–160
- Zhang, J., & Sclaroff, S. (2015). Exploiting surroundedness for saliency detection: a boolean map approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5), 889–902.
- Zhang, L., Tong, M. H., Marks, T. K., Shan, H., & Cottrell, G. W. (2008). Sun: A bayesian framework for saliency using natural statistics. *Journal of vision*, 8(7), 32–32.
- Zhang L, Zhang J, Lin Z, Lu H, He Y (2019) Capsal: Leveraging captioning to boost semantics for salient object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 6024–6033
- Zhang X, Wang T, Qi J, Lu H, Wang G (2018) Progressive attention guided recurrent network for salient object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 714–722
- Zhao, Q., & Koch, C. (2013). Learning saliency-based visual attention : A review. *Signal Processing*, 93(6), 1401–1407.
- Zhao R, Ouyang W, Li H, Wang X (2015) Saliency detection by multi-context deep learning. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 1265–1274
- Zhao T, Wu X (2019) Pyramid feature attention network for saliency detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 3085–3094
- Zhou, L., & Gu, X. (2020). Embedding topological features into convolutional neural network salient object detection. *Neural Networks*, 121, 308–318.
- Zhu W, Liang S, Wei Y, Sun J (2014) Saliency optimization from robust background detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp 2814–2821
- Zhuo, Y., Zhou, T. G., Rao, H. Y., Wang, J. J., Meng, M., Chen, M., et al. (2003). Contributions of the visual ventral pathway to long-range apparent motion. *Science*, 299(5605), 417–420.
- Zitnick CL, Dollár P (2014) Edge boxes: Locating object proposals from edges. In: *European conference on computer vision*, Springer, pp 391–405

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.