Computational Visual Media
https://doi.org/10.1007/s41095-019-0149-9 Vol. 5, No. 2, June 2019, 117–150
Review Article

Salient object detection: A survey

Ali Borji[1], Ming-Ming Cheng[2] (✉), Qibin Hou[2], Huaizu Jiang[3], and Jia Li[4]

Abstract Detecting and segmenting salient objects from natural scenes, often referred to as salient object detection, has attracted great interest in computer vision. While many models have been proposed and several applications have emerged, a deep understanding of achievements and issues remains lacking. We aim to provide a comprehensive review of recent progress in salient object detection and situate this field among other closely related areas such as generic scene segmentation, object proposal generation, and saliency for fixation prediction. Covering 228 publications, we survey i) roots, key concepts, and tasks, ii) core techniques and main modeling trends, and iii) datasets and evaluation metrics for salient object detection. We also discuss open problems such as evaluation metrics and dataset bias in model performance, and suggest future research directions.

Keywords salient object detection; saliency; visual attention; regions of interest

1 Introduction

Humans are able to detect visually distinctive, so called *salient*, scene regions effortlessly and rapidly in a *pre-attentive* stage. These filtered regions are then perceived and processed in finer detail for the extraction of richer high-level information, in an *attentive* stage. This capability has long been studied by cognitive scientists and has recently attracted much interest in the computer vision community, mainly because it helps to find the objects or regions that efficiently represent a scene, a useful step in complex vision problems such as scene understanding. Some topics that are closely or remotely related to visual saliency include: salient object detection [1], fixation prediction [2, 3], object importance [4–6], memorability [7], scene clutter [8], video interestingness [9–12], surprise [13], image quality assessment [14–16], scene typicality [17, 18], aesthetics [11], and scene attributes [19]. Given space limitations, this paper cannot fully explore all of the aforementioned research directions. Instead, we only focus on salient object detection, a research area that has greatly developed in the past twenty years, and in particular since 2007 [20].

1.1 What is salient object detection about?

Salient object detection or *salient object segmentation* is commonly interpreted in computer vision as a process that includes two stages: 1) *detecting the most salient object* and 2) *segmenting the accurate region of that object*. Rarely, however, models explicitly distinguish between these two stages (with few exceptions such as Refs. [21–23]). Following the seminal works by Itti et al. [24] and Liu et al. [25], models adopt the saliency concept to simultaneously perform the two stages together. This is witnessed by the fact that these stages have not been separately evaluated. Further, mostly area-based scores have been employed for model evaluation (e.g., precision–recall). The first stage does not necessarily need to be limited to only one object. The majority of existing models, however, attempt to segment the most salient object, although their prediction maps can be used to find several objects in a scene. The second stage falls into the realm of classic segmentation problems in computer vision but with the difference that here, accuracy is only determined by the most salient object.

1 MarkableAI, New York, USA. E-mail: ali@markable.ai.
2 TKLNDST, College of Computer Science, Nankai University, Tianjin, China. E-mail: cmm@nankai.edu.cn (✉).
3 University of Massachusetts Amherst, Amherst, MA, USA.
4 State Key Laboratory of Virtual Reality Technology and Systems, School of Computer Science and Engineering, Beihang University, Beijing, China. E-mail: jiali@buaa.edu.cn.
Manuscript received: 2019-05-22; accepted: 2019-05-27

In general, it is agreed that for good saliency detection a model should meet at least the following three criteria: 1) *good detection:* the probability of missing real salient regions and falsely marking the background as a salient region should be low, 2) *high resolution:* saliency maps should have high or full resolution to accurately locate salient objects and retain original image information, and 3) *computational efficiency:* as front-ends to other complex processes, these models should detect salient regions quickly.

1.2 Situating salient object detection

Salient object detection models usually aim to detect only the most salient objects in a scene and segment the whole extent of those objects. Fixation prediction models, on the other hand, typically try to predict where humans look, i.e., a small set of fixation points [31, 32]. Since both types of method output a single continuous-valued saliency map, where a higher value in this map indicates that the corresponding image pixel is more likely to be looked at, they can be used interchangeably.

A strong correlation exists between fixation locations and salient objects. Furthermore, humans often agree with each other when asked to choose the most salient object in a scene [22, 23, 26]. See Fig. 1.

Unlike salient object detection and fixation prediction models, object proposal models aim at producing a small set, typically a few hundreds or thousands, of overlapping candidate object bounding boxes or region proposals [33]. Object proposal generation and salient object detection are highly related. Saliency estimation is explicitly used as a cue in objectness methods [34, 35].

Image segmentation, also called semantic scene labeling or semantic segmentation, is one of the very well researched areas in computer vision (e.g., Ref. [36]). In contrast to salient object detection, where the output is a binary map, these models aim to assign a label, one out of several classes such as sky, road, and building, to each image pixel.

Figure 2 illustrates the differences between these research themes.

1.3 History of salient object detection

One of the earliest saliency models, proposed by Itti et al. [24], generated the *first wave* of interest across multiple disciplines including cognitive psychology, neuroscience, and computer vision. This model is an implementation of earlier general computational frameworks and psychological theories of bottom–up attention based on center–surround mechanisms (e.g., *feature integration theory* by Treisman and Gelade [50], the *guided search model* by Wolfe et al. [51], and the *computational attention architecture* by Koch and Ullman [52]). In Ref. [24], Itti et al. show some examples where their model is able to detect spatial discontinuities in scenes. Subsequent behavioral (e.g., Ref. [53]) and computational (e.g., Ref. [54]) investigations used fixations as a means to verify the saliency hypothesis and to compare models.

A *second wave* of interest surged with the works of Liu et al. [25, 55] and Achanta et al. [56] who defined saliency detection as a binary segmentation problem. These authors were inspired by some earlier models striving to detect salient regions or proto-objects (e.g., Ma and Zhang [57], Liu and Gleicher [58], and

Fig. 1 An example image in Borji et al. 's experiment [26] along with annotated salient objects. Dots represent 3-second free-viewing fixations.

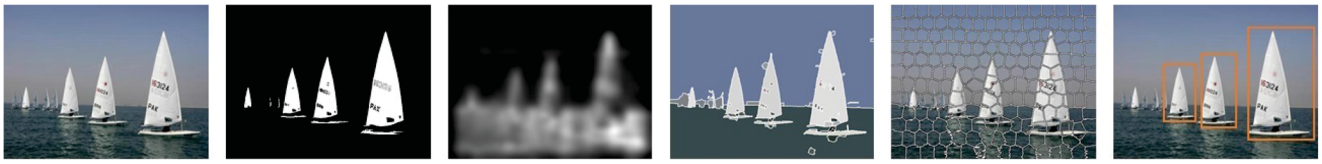


Fig. 2 Sample results produced by different models. Left to right: input image, salient object detection [27], fixation prediction [24], image segmentation (regions with various sizes) [28], image segmentation (superpixels with comparable sizes) [29], and object proposals (true positives) [30].

Walther and Koch [59]). A plethora of saliency models has emerged since then. It has been, however, less clear how this new definition relates to other established computer vision areas such as image segmentation (e.g., Refs. [60, 61]), category independent object proposal generation (e.g., Refs. [30, 34, 62]), fixation prediction (e.g., Refs. [54, 63–66]), and object detection (e.g., Refs. [67, 68]).

A *third wave* of interest has appeared recently with the surge in popularity of convolutional neural networks (CNNs) [69], and in particular with the introduction of fully convolutional neural networks [70]. Unlike the majority of classic methods based on contrast cues [1], CNN-based methods both eliminate the need for hand-crafted features, and alleviate the dependency on center bias knowledge, and hence have been adopted by many researchers. A CNN-based model normally contains hundreds of thousands of tunable parameters and neurons with variable receptive field sizes. Neurons with large receptive fields provide global information that can help better identify the most salient region in an image, while neurons with small receptive fields provide local information that can be leveraged to refine saliency maps produced by the higher layers. This allows highlighting salient regions and refining their boundaries. These desirable properties enable CNN-based models to achieve unprecedented performance compared to hand-crafted feature-based models. CNN models are gradually becoming the mainstream direction in salient object detection.

2 Survey of the state-of-the-art

In this section, we review related works in 3 categories, including: 1) salient object detection models, 2) applications, and 3) datasets. The similarity of various models means that it is sometimes hard to draw sharp boundaries between them. Here we mainly focus on the models contributing to the major waves in the chronicle shown in Fig. 3.

2.1 Old testament: classic models

A large number of approaches have been proposed for detecting salient objects in images in the past two decades. Except for a few models which attempt to segment objects-of-interest (e.g., Refs. [71–73]), most approaches aim to identify salient subsets from images first (i.e., compute a saliency map) and then integrate them to segment the entire salient object.

Visual subsets could be pixels, blocks, superpixels, or regions. Blocks are rectangular patches uniformly sampled from the image; pixels are 1×1 blocks. A superpixel or a region is a perceptually homogeneous image patch that is confined within intensity edges. Superpixels, in the same image, often have comparable but different sizes, while the shapes and sizes of regions may change remarkably. In this review, the term *block* is used to represent pixels and patches, while *superpixel* and *region* are used interchangeably.

In general, classic approaches can be categorized in two different ways depending on the type operation or attributes they exploit.

1. **Block-based versus region-based analysis**. Two types of visual subsets have been utilized: blocks and regions, to detect salient objects. Blocks were primarily adopted by early approaches, while regions became popular with the introduction of superpixel algorithms.

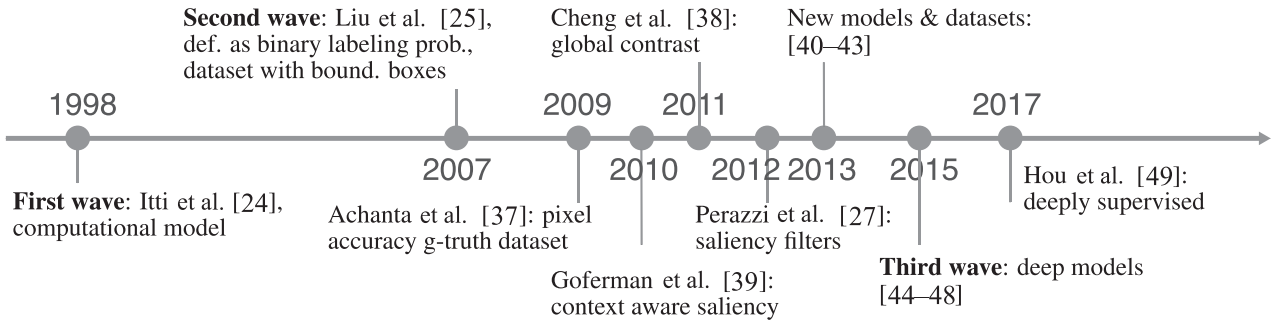


Fig. 3 A simplified chronicle of salient object detection modeling. The first wave started with the Itti et al. model [24], followed by the second wave with the introduction of the approach of Liu et al. [25] who were the first to define saliency as a binary segmentation problem. The third wave started with the surge of deep learning models and the model of Li and Yu [47].

2. **Intrinsic cues versus extrinsic cues**. A key step in detecting salient objects is to distinguish them from distractors. To do so, some approaches extract various cues only from the input image itself, to highlight targets and to suppress distractors (i.e., the intrinsic cues). However, other approaches argue that intrinsic cues are often insufficient to distinguish targets and distractors, especially when they share common visual attributes. To overcome this issue, they incorporate extrinsic cues such as user annotation, depth maps, or statistical information about similar images to facilitate detection of salient objects in the image.

Using the above model categorization, four combinations are thus possible. To structure our review, we group the models into three major subgroups: 1) *block-based models with intrinsic cues*, 2) *region-based models with intrinsic cues*, and 3) *models with extrinsic cues (both block- and region-based)*. Some approaches that do not easily fit into these subgroups are discussed in an *other classic models* subgroup. Reviewed models are listed in Table 1 (intrinsic models), Table 2 (extrinsic models), and Table 3 (other classic models).

2.1.1 Block-based models with intrinsic cues

In this subsection, we mainly review salient object detection models which utilize intrinsic cues extracted

Table 1 Salient object detection models with intrinsic cues (sorted by year). Elements: {PI = pixel, PA = patch, RE = region}, where prefixes m and h indicate multi-scale and hierarchical versions, respectively. Hypothesis: {CP = center prior, G = global contrast, L = local contrast, D = edge density, B = background prior, F = focus prior, O = objectness prior, CV = convexity prior, CS = center-surround contrast, CLP = color prior, SD = spatial distribution, BC = boundary connectivity prior, SPS = sparse noise}. Aggregation/optimization: {LN = linear, NL = non-linear, AD = adaptive, HI = hierarchical, BA = Bayesian, GMRF = Gaussian MRF, EM = energy minimization, and LS = least-square solver}. Code: {M= Matlab, C= C/C++, NA = not available, EXE = executable}

#	Model	Pub	Year	Elements	Hypothesis		Aggregation (optimization)	Code
					Uniqueness	Prior		
1	**FG** [57]	MM	2003	PI	L	—	—	NA
2	**RSA** [74]	MM	2005	PA	G	—	—	NA
3	**RE** [58]	ICME	2006	mPI+RE	L	—	LN	NA
4	**RU** [83]	TMM	2007	RE	—	P	LN	NA
5	**AC** [56]	ICVS	2008	mPA	L	—	LN	NA
6	**FT** [37]	CVPR	2009	PI	CS	—	—	C
7	**ICC** [77]	ICCV	2009	PI	L	—	LN	NA
8	**EDS** [76]	PR	2009	PI	—	ED	—	NA
9	**CSM** [90]	MM	2010	PI+PA	L	SD	—	NA
10	**RC** [84]	CVPR	2011	RE	G	—	—	C
11	**HC** [84]	CVPR	2011	RE	G	—	—	C
12	**CC** [91]	ICCV	2011	mRE	—	CV	—	NA
13	**CSD** [78]	ICCV	2011	mPA	CS	—	LN	NA
14	**SVO** [92]	ICCV	2011	PA+RE	CS	O	EM	M+C
15	**CB** [93]	BMVC	2011	mRE	L	CP	LN	M+C
16	**SF** [27]	CVPR	2012	RE	G	SD	NL	C
17	**ULR** [94]	CVPR	2012	RE	SPS	CP+CLP	—	M+C
18	**GS** [95]	ECCV	2012	PA/RE	—	B	—	NA
19	**LMLC** [96]	TIP	2013	RE	CS	—	BA	M+C
20	**HS** [42]	CVPR	2013	hRE	G	—	HI	EXE
21	**GMR** [97]	CVPR	2013	RE	—	B	—	M
22	**PISA** [89]	CVPR	2013	RE	G	SD+CP	NL	NA
23	**STD** [85]	CVPR	2013	RE	G	—	—	NA
24	**PCA** [80]	CVPR	2013	PA+PE	G	—	NL	M+C
25	**GU** [86]	ICCV	2013	RE	G	—	—	C
26	**GC** [86]	ICCV	2013	RE	G	SD	AD	C
27	**CHM** [79]	ICCV	2013	PA+mRE	CS+L	—	LN	M+C
28	**DSR** [98]	ICCV	2013	mRE	—	B	BA	M+C
29	**MC** [99]	ICCV	2013	RE	—	B	—	M+C
30	**UFO** [100]	ICCV	2013	RE	G	F+O	NL	M+C
31	**CIO** [101]	ICCV	2013	RE	G	O	GMRF	NA
32	**SLMR** [102]	BMVC	2013	RE	SPS	BC	—	NA
33	**LSMD** [103]	AAAI	2013	RE	SPS	CP+CLP	—	NA
34	**SUB** [87]	CVPR	2013	RE	G	CP+CLP+SD	—	NA
35	**PDE** [104]	CVPR	2014	RE	—	CP+B+CLP	—	NA
36	**RBD** [105]	CVPR	2014	RE	—	BC	LS	M

Table 2 Salient object detection models with extrinsic cues grouped by their adopted cues. For cues: {GT = ground-truth annotation, SI = similar images, TC = temporal cues, SCO = saliency co-occurrence, DP = depth, and LF = light field}. For saliency hypothesis: {P = generic properties, PRA = pre-attention cues, HD = discriminativity in high-dimensional feature space, SS = saliency similarity, CMP = complement of saliency cues, SP = sampling probability, MCO = motion coherence, RP = repeatedness, RS = region similarity, C = corresponding, and DK = domain knowledge}. Others: {CRF = conditional random field, SVM = support vector machine, BDT = boosted decision tree, and RF = random forest}

#	Model	Pub	Year	Cues	Elements	Hypothesis		Aggregation (optimization)	GT form	Code
						Uniqueness	Prior			
1	**LTD** [25]	CVPR	2007	GT	mPI+PA+RE	L+CS	SD	CRF	BB	NA
2	**OID** [109]	ECCV	2010	GT	mPI+PA+RE	L+CS	SD	mixtureSVM	BB	NA
3	**LGCR** [110]	BMVC	2010	GT	RE	—	P	BDT	BM	NA
4	**DRFI** [40]	CVPR	2013	GT	mRE	L	B+P	RF	BM	M+C
5	**LOS** [111]	CVPR	2014	GT	RE	L+G	PRA+B+SD+CP	SVM	BM	NA
6	**HDCT** [112]	CVPR	2014	GT	RE	L+G	SD+P+HD	BDT+LS	BM	M

#	Model	Pub	Year	Cues	Elements	Hypothesis		Aggregation (optimization)	GT necessity	Code
						Uniqueness	Prior			
7	**VSIT** [113]	ICCV	2009	SI	PA	—	SS	—	yes	NA
8	**FIEC** [114]	CVPR	2011	SI	PI+PA	L	—	LN	no	NA
9	**SA** [115]	CVPR	2013	SI	PI	—	CMP	CRF	yes	NA
10	**LBI** [35]	CVPR	2013	SI	PA	SP	—	—	no	M+C

#	Model	Pub	Year	Cues	Elements	Hypothesis		Aggregation (optimization)	Type	Code
						Uniqueness	Prior			
11	**LC** [116]	MM	2006	TC	PI+PA	L	—	LN	online	NA
12	**VA** [117]	ICPR	2008	TC	mPI+PA+RE	L	CS+SD+MCO	CRF	offline	NA
13	**SEG** [108]	ECCV	2010	TC	PA+PI	CS	MCO	CRF	offline	M+C
14	**RDC** [118]	CSVT	2013	TC	RE	L	—	—	offline	NA

#	Model	Pub	Year	Cues	Elements	Hypothesis		Aggregation (optimization)	Image number	Code
						Uniqueness	Prior			
15	**CSIP** [119]	TIP	2011	SCO	mRE	—	RS	LN	two	M+C
16	**CO** [120]	CVPR	2011	SCO	PI+PA	G	RP	—	multiple	NA
17	**CBCO** [121]	TIP	2013	SCO	RE	G	SD+C	NL	multiple	NA

#	Model	Pub	Year	Cues	Elements	Hypothesis		Aggregation (optimization)	Source	Code
						Uniqueness	Prior			
18	**LS** [122]	CVPR	2012	DP	RE	G	DK	NL	stereo images	NA
19	**DRM** [123]	BMVC	2013	DP	RE	G	—	SVM	Kinect	NA
20	**SDLF** [107]	CVPR	2014	LF	mRE	G	F+B+O	NL	Lytro camera	NA

Table 3 Other salient object detection models

#	Model	Pub	Year	Type	Code
1	**COMP** [128]	ICCV	2011	Localization	NA
2	**GSAL** [129]	CVPR	2012	Localization	NA
3	**CTXT** [130]	ICCV	2011	Segmentation	NA
4	**LCSP** [131]	IJCV	2014	Segmentation	NA
5	**BENCH** [132]	ECCV	2012	Aggregation	M
6	**SIO** [133]	SPL	2013	Optimization	NA
7	**ACT** [21]	PAMI	2012	Active	C
8	**SCRT** [22]	CVPR	2014	Active	NA
9	**WISO** [23]	TIP	2014	Active	NA

from blocks. Following the seminal work of Itti et al. [24], salient object detection is widely defined as capturing uniqueness, distinctiveness, or rarity in a scene.

In early works [56–58], uniqueness was often computed as the pixel-wise center–surround contrast. Hu et al. [74] represent the input image in a 2D space using the polar transformation of its features. Each region in the image is then mapped into a 1D linear subspace. Afterwards, generalized principal component analysis (GPCA) [75] is used to estimate the linear subspaces without actually segmenting the image. Finally, salient regions are selected by measuring feature contrast and geometric properties of regions. Rosin [76] proposes an efficient approach for detecting salient objects. His approach is parameter-free and requires only very simple pixel-wise operations such as edge detection, threshold decomposition, and moment preserving binarization. Valenti et al. [77] propose an isophote-based framework where the saliency map is estimated by linearly combining saliency maps computed in terms of curvedness, color boosting, and isocenter clustering.

In an influential study, Achanta et al. [37] adopt a

frequency-tuned approach to compute full resolution saliency maps. The saliency of pixel x is computed as

$$s(x) = \| I_\mu - I_{\omega_{hc}}(x) \|^2 \quad (1)$$

where I_μ is the mean pixel value of the image (e.g., RGB/Lab features) and $I_{\omega_{hc}}$ is a Gaussian blurred version of the input image (e.g., using a 5×5 kernel).

Without prior knowledge of the sizes of salient objects, multi-scale contrast is frequently adopted for robustness [25, 58]. An L-layer Gaussian pyramid is first constructed (as in Refs. [25, 58]). The saliency score of pixel x in the image at the lth level of this pyramid (denoted as $I^{(l)}$) is defined as

$$s(x) = \sum_{l=1}^{L} \sum_{x' \in \mathcal{N}(x)} ||I^{(l)}(x) - I^{(l)}(x')||^2 \quad (2)$$

where $\mathcal{N}(x)$ is a neighborhood window centered at x (e.g., 9×9 pixels). Even with such multi-scale enhancement, intrinsic cues derived at pixel level are often too poor to support object segmentation. To address this, some works (e.g., Refs. [25, 56, 78, 79]) extended contrast analysis to the patch level (comparing patches to their neighbors).

Later in Ref. [78], Klein and Frintrop proposed an information-theoretic approach to compute center–surround contrast using the Kullback–Leibler divergence between distributions of features such as intensity, color, and orientation. Li et al. [79] formulated center–surround contrast as a cost-sensitive max-margin classification problem. The center patch is labeled as a positive sample while the surrounding patches are all used as negative samples. The saliency of the center patch is then determined by its separability from surrounding patches based on a trained cost-sensitive support vector machine (SVM).

Some works have defined patch uniqueness as a patch's global contrast to other patches [39]. Intuitively, a patch is considered to be salient if it is significantly different from the other patches most similar to it; their spatial distances are taken into account. Similarly, Borji and Itti computed local and global patch rarity in RGB and Lab color spaces and fused them to predict fixation locations [65]. In recent work [80], Margolin et al. define the uniqueness of a patch by measuring its distance to the average patch based on the observation that distinctive patches are more scattered than non-distinctive ones in the high-dimensional space. To further incorporate the patch distributions, the uniqueness of a patch is measured by projecting its path to the average patch onto the principal components of the image.

To sum up, approaches in this section aim to detect salient objects based on pixels or patches utilizing only intrinsic cues. These approaches usually suffer from two shortcomings: 1) high-contrast edges usually stand out instead of the salient object, and 2) the boundary of the salient object is not preserved well (especially when using large blocks). To overcome these issues, some methods propose to compute saliency based on regions. This offers two main advantages. First, the number of regions is far fewer than the number of blocks, offering the potential to develop highly efficient and fast algorithms. Second, more informative features can be extracted from regions, leading to better performance. Such region-based approaches are discussed in the next subsection.

2.1.2 Region-based models with intrinsic cues

Saliency models in the second subgroup adopt intrinsic cues extracted from image *regions* generated using methods such as graph-based segmentation [81], mean-shift [28], SLIC [29], or Turbopixels [82]. Unlike block-based models, region-based models often segment an input image into regions aligned with intensity edges first, and then compute a regional saliency map.

As an early attempt, in Ref. [58], regional saliency score is defined as the average saliency score of the region's pixels, defined in terms of multi-scale contrast. Yu and Wong [83] propose a set of rules to determine the background scores of each region based on observations from background and salient regions. Saliency, defined as uniqueness in terms of *global regional contrast*, is widely studied in many approaches [42, 84–87]. In Ref. [84], a region-based saliency algorithm is introduced by measuring the global contrast between the target region and all other image regions. In a nutshell, an image is first segmented into N regions $\{r_i\}_{i=1}^N$. Saliency of region r_i is measured as

$$s(r_i) = \sum_{j=1}^{N} w_{ij} D_r(r_i, r_j) \quad (3)$$

where $D_r(r_i, r_j)$ captures the appearance contrast between two regions. Higher saliency scores are

assigned to regions with large global contrast. w_{ij} is a weight linking regions r_i and r_j, which incorporates spatial distance and region size. Perazzi et al. [27] demonstrate that if $D_r(r_i, r_j)$ is defined as the Euclidean color distance between r_i and r_j, global contrast can be computed using efficient filtering based techniques [88].

In addition to color uniqueness, distinctiveness of complementary cues such as texture [85] and structure [89] are also considered for salient object detection. Margolin et al. [80] propose to combine regional uniqueness and patch distinctiveness to form a saliency map. Instead of maintaining a hard region index for each pixel, a soft abstraction is proposed in Ref. [86] to generate a set of large-scale perceptually homogeneous regions using histogram quantization and Gaussian mixture models (GMMs). By avoiding hard decisions about boundaries of superpixels, such soft abstraction provides large spatial support which results in a more uniform saliency region.

In Refs. [93], Jiang et al. propose a *multi-scale local region contrast* based approach, which calculates saliency values across multiple segmentations for robustness purposes and combines these regional saliency values to obtain a pixel-wise saliency map. A similar idea for estimating regional saliency using multiple hierarchical segmentations is adopted in Refs. [42, 98]. Li et al. [79] extend pairwise local contrast by building a hypergraph, constructed by non-parametric multi-scale clustering of superpixels, to capture both internal consistency and external separation of regions. Salient object detection is then cast as finding salient vertices and hyperedges in the hypergraph.

Salient objects, in terms of uniqueness, can also be defined as *sparse noise* in a certain feature space in which the input image is represented as a low-rank matrix [94, 102, 103]. The basic assumption is that non-salient regions (i.e., background) can be explained by the low-rank matrix while the salient regions are indicated by sparse noise.

Based on such a general low-rank matrix recovery framework, Shen and Wu [94] propose a unified approach to incorporate traditional low-level features with higher-level guidance, e.g., *center prior*, *face prior*, and *color prior*, to detect salient objects based on a learned feature transformation. (Although extrinsic ground-truth annotations are adopted to learn high-level priors and the feature transformation, we classify this model with intrinsic models to better organize the low-rank matrix recovery based approaches. Additionally, we treat face and color priors as universal intrinsic cues for salient object detection). Instead, Zou et al. [102] propose to exploit bottom–up segmentation as a guidance cue for low-rank matrix recovery, for robustness. Similar to Ref. [94], high-level priors are also adopted in Ref. [103], where tree-structured sparsity-inducing norm regularization is introduced to hierarchically describe the image structure, in order to uniformly highlight the entire salient object.

In addition to capturing uniqueness, more and more priors have also been proposed for salient object detection. The *spatial distribution prior* [25] implies that the more widely a color is distributed in the image, the less likely a salient object is to contain this color. The spatial distribution of superpixels can also be efficiently evaluated in linear time using the Gaussian blurring kernel, in a similar way to computing global regional contrast in Eq. (3). Such a spatial distribution prior is also considered in Ref. [89], and is evaluated in terms of both color and structural cues.

A center prior assumes that a salient object is more likely to be found near the image center, and that the background tends to be far away from the image center. To this end, the *backgroundness prior* is adopted for salient object detection in Refs. [95, 97–99], assuming that a narrow border of the image forms the background region, i.e., the pseudo-background. With this pseudo-background as a reference, regional saliency can be computed as the contrast of regions versus "background". In Ref. [97], a two-stage saliency computation framework is proposed based on manifold ranking on an undirected weighted graph. In the first stage, regional saliency scores are computed based on the relevance given to each side of the pseudo-background queries. In the second stage, the saliency scores are refined based on the relevance given to the initial foreground. In Ref. [98], saliency computation is formulated in terms of dense and sparse reconstruction errors with respect to the pseudo-background. The dense reconstruction error of each region is computed from principal component analysis (PCA) of the background templates, while the sparse reconstruction error is defined as the

residual after sparse representation of the background templates. These two types of reconstruction errors are propagated to pixels in multiple segmentations, which are fused to form the final saliency map. Jiang et al. [99] formulate saliency detection via absorbing Markov chains, in which the transient and absorbing nodes are superpixels around the image center and border respectively. The saliency of each superpixel is computed as the absorption time between the transient node and the absorbing nodes of the Markov chain.

Beyond these approaches, the generic *objectness prior* is also used to facilitate salient object detection by leveraging object proposals [34]. Although it is learned from training data, we also tend to treat it as a universal intrinsic cue for salient object detection. Chang et al. [92] present a computational framework by fusing the objectness and regional saliency into a graphical model. These two terms are jointly estimated by iteratively minimizing an energy function that encodes their mutual interaction. In Ref. [100], region objectness is defined as the average objectness values of the pixels within the region; it is incorporated into regional saliency computation. Jia and Han [101] compute the saliency of each region by comparing it to the "soft" foreground and background according to the objectness prior.

Salient object detection relying on the pseudo-background assumption may fail sometimes, especially when the object touches the image border. To overcome this problem, a *boundary connectivity prior* is utilized in Refs. [84, 105]. Intuitively, salient objects are much less connected to the image border than objects in the background are. Thus, the boundary connectivity score of a region can be estimated according to the ratio of its length along the image border to the spanning area of this region [105]. The latter can be computed based on the region's geodesic distances to the pseudo-background and other regions respectively. Such a boundary connectivity score is integrated into a quadratic objective function to get the final optimized saliency map. It is worth noting that similar ideas of boundary connectivity prior are also investigated in [102] as *segmentation prior* and as *surroundingness* in Ref. [106].

The *focus prior*, the fact that a salient object is often photographed in focus to attract more attention, has been investigated in Refs. [100, 107]. Jiang et al. [100] calculate the focus score from the degree of focal blur. By modeling defocusing as the convolution of a sharp image with a point spread function, approximated by a Gaussian kernel, the pixel-level degree of focus can be estimated as the standard deviation of the Gaussian kernel by scale space analysis. A regional focus score is computed by propagating the focus score and/or sharpness at the boundary and interior edge pixels. The saliency score is finally derived from a non-linear combination of uniqueness (global contrast), objectness, and focus scores.

Performance of salient object detection based on regions can be affected by choice of segmentation parameters. In addition to other approaches based on multi-scale regions [42, 79, 93], single-scale potential salient regions are extracted by solving the facility location problem in Ref. [87]. An input image is first represented as an undirected graph of superpixels, where a much smaller set of candidate region centers is then generated through agglomerative clustering. On this set, a submodular objective function is built to maximize the similarity. By applying a greedy algorithm, the objective function can be iteratively optimized to group superpixels into regions whose saliency values are further measured via the regional global contrast and spatial distribution.

The Bayesian framework can also be exploited for saliency computation [96, 108], formulated as estimating the posterior probability of pixel x being foreground given the input image I. To estimate the saliency prior, a convex hull H is first estimated around the detected points of interest. The convex hull H, which divides the image I into the inner region R_{I} and outer region R_{O}, provides a coarse estimation of foreground as well as background, and can be adopted for likelihood computation. Liu et al. [104] use an optimization-based framework for detecting salient objects. As in Ref. [96], a convex hull is roughly estimated to partition an image into pure background and potential foreground. Then, saliency seeds are learned from the image, while a guidance map is learned from background regions, as well as human prior knowledge. Using these cues, a general linear elliptic system with Dirichlet boundary is introduced to model diffusion from seeds to other regions to generate a saliency map.

Among the models reviewed in this subsection, there are three main types of region adopted for saliency computation. Irregular regions of varying sizes can be generated using a graph-based segmentation algorithm [81], mean-shift algorithm [28], or clustering (quantization). On the other hand, with recent progress in superpixel algorithms, compact regions with comparable sizes are also popular choices, using the SLIC algorithm [29], Turbopixel algorithm [82], etc. The main difference between these two types of regions is whether the influence of region size should be taken into account. Furthermore, soft regions can also be considered for saliency analysis, where each pixel maintains a probability of belonging to each region (component) instead of having a hard region label (e.g., fitted by a GMM). To further enhance robustness of segmentation, regions can be generated based on multiple segmentations or in a hierarchical way. Generally, single-scale segmentation is faster, while multi-scale segmentation can improve the overall quality of results.

To measure the saliency of regions, uniqueness, usually in the form of global and local regional contrast, is still the most frequently used feature. In addition, more and more complementary priors for regional saliency have been investigated to improve the overall results, such as backgroundness, objectness, focus, and boundary connectivity. Compared to block-based saliency models, incorporation of these priors is the main advantage of region-based saliency models. Furthermore, regions provide more sophisticated cues (e.g., a color histogram) to better capture the salient object in a scene, in contrast to pixels and patches. Another benefit of defining saliency using regions is related to efficiency. Since the number of regions in an image is far fewer than the number of pixels, computing saliency at region level can significantly reduce the computational cost while producing full-resolution saliency maps.

Notice that the approaches discussed in this subsection only utilize intrinsic cues. In the next subsection, we review how to incorporate extrinsic cues to facilitate the detection of salient objects.

2.1.3 Models with extrinsic cues

Models in the third subgroup adopt *extrinsic cues* to assist in the detection of salient objects in images and videos. In addition to those visual cues observed in the single input image, extrinsic cues can be derived from ground-truth annotation of training images, similar images, video sequences, a set of input images containing the common salient objects, depth maps, or light field images. In this section, we will review such models according to the type of extrinsic cues used. Table 2 lists all models with extrinsic cues; each method is highlighted with several predefined attributes.

Salient object detection with similar images. With the availability of an increasingly large amount of visual content on the web, salient object detection by leveraging visually similar images to the input image has been studied in recent years. Generally, given the input image I, K similar images $\mathcal{C}_I = \{I_k\}_{k=1}^K$ are first retrieved from a large collection of images $\mathcal{C}$. Salient object detection in the input I can be assisted by examining these similar images.

In some studies, it is assumed that saliency annotations of $\mathcal{C}$ are available. For example, Marchesotti et al. [113] propose to describe each indexed image I_k by a pair of descriptors $(\boldsymbol{f}_{I_k}^+, \boldsymbol{f}_{I_k}^-)$, which respectively denote the feature descriptors (Fisher vector) of the salient and non-salient regions according to the saliency annotations. To compute the saliency map, each patch p_x of the input image is described by a Fisher vector $\boldsymbol{f}_x$. Saliencies of patches are computed according to their contrast with foreground and background region features: $\{(\boldsymbol{f}_{I_k}^+, \boldsymbol{f}_{I_k}^-)\}_{k=1}^K$.

Alternatively, based on the observation that different features contribute differently to the saliency analysis of each image, Mai et al. [115] propose to learn image specific rather than universal weights to fuse the saliency maps computed on different feature channels. To this end, the CRF aggregation model of saliency maps is trained only on retrieved similar images to account for the dependence of aggregation on individual images. We will give further technical details of Ref. [115] in Section 2.1.4.

Saliency based on similar images works well if large-scale image collections are available. Saliency annotation, however, is time consuming, tedious, and even intractable on such collections. To mitigate this, some methods leverage *unannotated* similar images. Using web-scale image collections $\mathcal{C}$, Wang et al. [114] propose a simple yet effective saliency estimation

algorithm. The pixel-wise saliency map is computed as

$$s(x) = \sum_{k=1}^{K} ||I(x) - \tilde{I}_k(x)||_1 \quad (4)$$

where $\tilde{I}_k$ is a geometrically warped version of I_k with reference I. The main insight is that similar images offer good approximations to the background regions while salient regions might not be well-approximated.

Siva et al. [35] propose a probabilistic formulation for saliency computation as a sampling problem. A patch p_x is considered to be salient if it has the low probability of being sampled from the images $\mathcal{C}_I \cup I$. In other words, a high saliency score will be given to p_x if it is unusual among a bag of patches extracted from similar images.

Co-saliency object detection. Instead of concentrating on computing saliency in a single image, co-salient object detection algorithms focus on discovering *common* salient objects shared by multiple input images $\{I^i\}_{i=1}^M$. Such objects can be the same object from different viewpoints, or objects in the same category, sharing similar visual appearance. Note that the key characteristic of co-salient object detection algorithms is that their input is *a set* of images, while classical salient object detection models only need a *single* input image.

Co-saliency detection is closely related to the concept of image co-segmentation, which aims to segment similar objects from multiple images [124, 125]. As stated in Ref. [121], three major differences exist between co-saliency and co-segmentation. First, co-saliency detection algorithms only focus on detecting common salient objects, while similar but non-salient background might be also segmented out in co-segmentation approaches [126, 127]. Second, some co-segmentation methods, e.g., Ref. [125], need user input to guide the segmentation process in ambiguous situations. Third, salient object detection often serves as a pre-processing step, and thus more efficient algorithms are preferred than for co-segmentation algorithms, especially when processing a large number of images.

Li and Ngan [119] propose a method to compute co-saliency for an image pair with some objects in common. The co-saliency is defined as the inter-image correspondence, i.e., low saliency values should be given to dissimilar regions. Similarly in Ref. [120], Chang et al. propose to compute co-saliency by exploiting the additional *repeatedness* property across multiple images. Specifically, the co-saliency score of a pixel is defined as the multiplication of its traditional saliency score [39] and its repeatedness likelihood over the input images. Fu et al. [121] propose a cluster-based co-saliency detection algorithm by exploiting the well-established global contrast and spatial distribution concepts on a single image. Additionally, corresponding cues over multiple images are introduced to account for saliency co-occurrence.

2.1.4 Other classic models

In this section, we review algorithms that aim to directly segment or localize salient objects with bounding boxes, and algorithms that are closely related to saliency detection. Some subsections offer a different categorization of some models covered in the previous sections (e.g., supervised versus unsupervised). See Table 3.

Localization models. Liu et al. [25] convert the binary segmentation map to bounding boxes. The final output is a set of rectangles around salient objects. Feng et al. [128] define saliency for a sliding window as its composition cost using the remaining image parts. Based on an over-segmentation of the image, the local maxima, which can efficiently be found among all sliding windows in a brute-force manner, are assumed to correspond to salient objects.

The basic assumption in many previous approaches is that at least one salient object exists in the input image. This may not always hold as some *background images* contain no salient objects at all. In Ref. [129], Wang et al. investigate the problem of localizing and predicting the *existence* of salient objects in thumbnail images. Specifically, each image is described by a set of features extracted in multiple channels. The existence of salient objects is formulated as a binary classification problem. For localization, a regression function is learned using random forest regression on training samples to directly output the position of the salient object.

Segmentation models. Segmenting salient objects is closely related to the figure-ground problem, which is essentially a binary classification problem, trying to separate the salient object from the background. Yu et al. [90] utilize the complementary characteristics of imperfect saliency maps generated by different contrast-based saliency models. Specifically, two complementary saliency maps are first generated

for each image, including a sketch-like map and an envelope-like map. The sketch-like map can accurately locate parts of the most salient object (i.e., skeleton with high precision), while the envelope-like map can roughly cover the entire salient object (i.e., envelope with high recall). With these two maps, reliable foreground and background regions can be detected in each image by first training a pixel classifier. By labeling all other pixels with this classifier, salient object can be detected as a whole. This method is extended in Ref. [131] by learning complementary saliency maps for the purpose of salient object segmentation.

Lu et al. [91] exploit the convexity (concavity) prior for salient object segmentation. This prior assumes that the region on the convex side of a curved boundary tends to belong to the foreground. Based on this assumption, concave arcs are first found on the contours of superpixels. The convexity context of a concave arc is defined by windows close to the arc. An undirected weight graph is then built over superpixels with concave arcs, where the weights between vertices are determined by summing the concavity context at different scales in the hierarchical segmentation of the image. Finally, the normalized cut algorithm [134] is used to separate the salient object from the background.

To leverage contextual cues more effectively, Wang et al. [130] propose to integrate an auto-context classifier [135] into an iterative energy minimization framework to automatically segment the salient object. The auto-context model is a multi-layer boosting classifier on each pixel and its surroundings to predict whether it is associated with the target concept. The subsequent layer is built on the classification of the previous layer. Hence, through the layered learning process, spatial context is automatically utilized for more accurate segmentation of the salient object.

Supervised versus unsupervised models. The majority of existing learning-based works on saliency detection focus on the supervised scenario, i.e., learning a salient object detector given a set of training samples with ground-truth annotation. The aim here is to separate salient elements from background elements.

Each element (e.g., a pixel or a region) in the input image is represented by a feature vector $\boldsymbol{f} \in \mathbb{R}^D$, where D is the feature dimension. Such a feature vector is then mapped to a saliency score $s \in \mathbb{R}^+$ based on the learned linear or non-linear mapping function $f : \mathbb{R}^D \to \mathbb{R}^+$.

One can assume the mapping function f is linear, i.e., $s = \boldsymbol{w}^{\mathrm{T}}\boldsymbol{f}$, where $\boldsymbol{w}$ denotes the combination weights of all components in the feature vector. Liu et al. [25] learn the weights with a conditional random field (CRF) model trained on rectangular annotations of the salient objects. In recent work [111], the large-margin framework is adopted to learn the weights $\boldsymbol{w}$.

Due to the highly non-linear nature of the saliency mechanism, however, a linear mapping may not perfectly capture the characteristics of saliency. To this end, the linear approach is extended in Ref. [109], where a mixture of linear support vector machines (SVMs) is adopted to partition the feature space into a set of sub-regions that are linearly separable using a divide-and-conquer strategy. In each region, a linear SVM, its mixture weights, and the combination parameters of the saliency features are learned for better saliency estimation. Alternatively, other non-linear classifiers such as boosted decision trees (BDTs) [110, 112] and random forest (RFs) [40] may also be utilized.

Generally speaking, supervised approaches allow richer representations for the elements compared with heuristic methods. In seminal work on supervised salient object detection, Liu et al. [25] propose a set of features including local multi-scale contrast, regional center–surround histogram distance, and global color spatial distribution. As for models with only intrinsic cues, region-based representation for salient object detection has become increasingly popular as more sophisticated descriptors can be extracted at region level. Mehrani and Veksler [110] demonstrate promising results by considering generic regional properties, e.g., color and shape, which are widely used in other applications like image classification. Jiang et al. [40] propose a regional saliency descriptor including regional local contrast, regional backgroundness, and regional generic properties. In Refs. [111, 112], each region is described by a set of features such as local and global contrast, backgroundness, spatial distribution, and the center prior. Pre-attentive features are also considered in Ref. [111].

Usually, richer representations result in feature vectors with higher dimensions, e.g., $D = 93$ in Ref. [40] and $D = 75$ in Ref. [112]. With the availability of large collections of training samples, the learned classifier is capable of automatically integrating such richer features and selecting the most discriminative ones. Therefore, better performance can be expected than with heuristic methods.

Some models have utilized unsupervised techniques. In Ref. [35], saliency computation is formulated in a probabilistic framework as a sampling problem. The saliency of each image patch is proportional to its sampling probability from all patches extracted from both the input image and similar images retrieved from a corpus of unlabeled images. In Ref. [136], cellular automata are exploited for unsupervised salient object detection.

Aggregation and optimization models. Given M saliency maps $\{S_i\}_{i=1}^{M}$, coming from different salient object detection models or hierarchical segmentations of the input image, aggregation models try to form a more accurate saliency map. Let $S_i(x)$ denote the saliency value of pixel x in the ith saliency map. In Ref. [132], Borji et al. propose a standard saliency aggregation method as follows:

$$S(x) = P(s_x = 1|\boldsymbol{f}_x) \propto \frac{1}{Z}\sum_{i=1}^{M}\zeta(S_i(x)) \qquad (5)$$

where $\boldsymbol{f}_x = (S_1(x), \cdots, S_M(x))$ are the saliency scores for pixel x and $s_x = 1$ indicates x is labeled as salient. $\zeta(\cdot)$ is a real-valued function which takes the following form:

$$\zeta_1(z) = z, \quad \zeta_2(z) = \exp(z), \quad \zeta_3(z) = -\frac{1}{\log(z)} \qquad (6)$$

Inspired by the aggregation model in Ref. [132], Mai et al. [115] propose two aggregation solutions. The first solution adopts pixel-wise aggregation:

$$P(s_x = 1|\boldsymbol{f}_x; \lambda) = \sigma\left(\sum_{i=1}^{M}\lambda_i S_i(x) + \lambda_{M+1}\right) \qquad (7)$$

where $\lambda = \{\lambda_i | i = 1, \cdots, M+1\}$ is the set of model parameters and $\sigma(z) = 1/(1+\exp(-z))$. However, they note one potential problem of such direct aggregation, its ignorance of interactions between neighboring pixels. Inspired by Ref. [55], they propose the second solution which uses the CRF to aggregate saliency maps of multiple methods to capture the relation between neighboring pixels. The parameters of the CRF aggregation model are optimized on the training data. The saliency of each pixel is the posterior probability of being labeled as salient with the trained CRF.

Alternatively, Yan et al. [42] integrate saliency maps computed on hierarchical segmentations of the image into a tree-structured graphical model, where each node corresponds to a region in every level of the hierarchy. Thanks to the tree structure, saliency inference can efficiently be conducted using belief propagation. In fact, solving the three layer hierarchical model is equivalent to applying a weighted average to all single-layer maps. Unlike naive multi-layer fusion, this hierarchical inference algorithm can select optimal weights for each region instead of a global weighting.

Li et al. [133] propose to optimize the saliency values of all superpixels in an image to simultaneously meet several saliency criteria including visual rarity, center-bias, and mutual correlation. Based on the correlations (similarity scores) between region pairs, the saliency value of each superpixel is optimized by quadratic programming when considering the influences of all other superpixels. Let w_{ij} denote the correlation between two regions r_i and r_j. The saliency values $\{s_i\}_{i=1}^{N}$ (denoting $s(r_i)$ as s_i for short) can be optimized by solving:

$$\begin{aligned}\min_{\{s_i\}_{i=1}^{N}} &\sum_{i=1}^{N} s_i \sum_{j\neq i}^{N} w_{ij} + \lambda_{\mathrm{c}}\sum_{i=1}^{N} s_i \mathrm{e}^{d_i/d_{\mathrm{D}}} \\ &+ \lambda_{\mathrm{r}}\sum_{i=1}^{N}\sum_{j\neq i}^{N}(s_i - s_j)^2 w_{ij}\mathrm{e}^{-d_{ij}/d_{\mathrm{D}}} \\ \text{such that } \ 0 \leqslant s_i \leqslant 1, \forall i, \text{ and } &\sum_{i=1}^{N} s_i = 1 \end{aligned} \qquad (8)$$

Here d_{D} is half the image diagonal length, and d_{ij} and d_i are spatial distances from r_i to r_j and the image center, respectively. In the optimization, the saliency value of each superpixel is optimized by quadratic programming, considering the influences of all other superpixels. Zhu et al. [105] also adopt a similar optimization-based framework to integrate multiple foreground/background cues as well as smoothness terms to automatically infer optimal saliency values.

The Bayesian framework is adopted to more effectively integrate the complementary dense and sparse reconstruction errors [98]. A fully-connected Gaussian Markov random field between each pair of regions is constructed to enforce consistency between salient regions [101], which permitting

efficient computation of the final regional saliency scores.

Active models. Inspired by interactive segmentation models (e.g., Refs. [137, 138]), a new trend has emerged recently, explicitly decoupling the two stages of saliency detection mentioned in Section 1.1: 1) detecting the most salient object and 2) segmenting it. Some studies propose to perform active segmentation by utilizing the advantages of both fixation prediction and segmentation models. For example, Mishra et al. [21] combine multiple cues (e.g., color, intensity, texture, stereo, and/or motion) to predict fixations. The "optimal" closed contour for the salient object around the fixation point is then segmented in polar space. Li et al. [22] propose a model composed of two components: a *segmenter* that proposes candidate regions and a *selector* that gives each region a saliency score (using a fixation prediction model). Similarly, Borji [23] proposes to first roughly locate the salient object at the peak of the fixation map (or its estimation using a fixation prediction model) and then segment the object using superpixels. The last two algorithms adopt annotations to determine the upper-bound of segmentation performance, propose datasets with multiple objects in scenes, and provide new insight into the inherent connections between fixation prediction and salient object segmentation.

Salient object detection in video. In addition to spatial information, video sequences provide temporal cues, e.g., motion, which facilitates salient object detection. Zhai and Shah [116] first estimate keypoint correspondences between two consecutive frames. Motion contrast is computed based on planar motions (the homography) between images, which is estimated by applying RANSAC to point correspondences. Liu et al. [117] extend their spatial saliency features [25] to the motion field resulting from an optical flow algorithm. Using the colorized motion field as the input image, local multi-scale contrast, regional center-surround distance, and global spatial distribution are computed and finally integrated in a linear way. Rahtu et al. [108] integrate spatial saliency into an energy minimization framework by considering the temporal coherence constraint. Li et al. [118] extend regional contrast-based saliency to the spatio-temporal domain. Given an over-segmentation of the frames of the video sequence, spatial and temporal region matches between each two consecutive frames are estimated in a interactive manner on an undirected unweighted matching graph, based on the regions' colors, textures, and motion features. The saliency of a region is determined by computing its local contrast to the surrounding regions not only in the present frame but also in the temporal domain.

Salient object detection with depth. We live in a 3D environment in which stereoscopic content provides additional depth cues for guiding visual attention and understanding our surroundings. This point is further validated by Lang et al. [139] through experimental analysis of the importance of depth cues for eye fixation prediction. Recently, researchers have started to study how to exploit depth cues for salient object detection [122, 123]; these might be captured indirectly from stereo images or directly using a depth camera (e.g., Kinect).

The most straightforward extension is to adopt the widely used hypotheses introduced in Section 2.1.1 and 2.1.2 to the depth channel, e.g., global contrast on the depth map [122, 123]. Furthermore, Niu et al. [122] demonstrate how to leverage domain knowledge in stereoscopic photography to compute the saliency map. The input image is first segmented into regions $\{r_i\}$. In practice, the regions at the focus of attention are often assigned small or zero disparities to minimize the *vergence-accommodation conflict*. Therefore, the first type of regional saliency based on disparity is defined as

$$s_{d,1}(r_i) = \begin{cases} (d_{\max} - \bar{d}_i)/d_{\max}, & \bar{d}_i \geqslant 0 \\ (d_{\min} - \bar{d}_i)/d_{\min}, & \bar{d}_i < 0 \end{cases} \tag{9}$$

where $d_{\max}$ and $d_{\min}$ are the maximal and minimal disparities, respectively. $\bar{d}_i$ denotes the average disparity in region r_i. Additionally, objects with negative disparities are perceived as popping out of the scene. The second type of regional stereo saliency is then defined as

$$s_{d,2}(r_i) = \frac{d_{\max} - \bar{d}_i}{d_{\max} - d_{\min}} \tag{10}$$

Stereo saliency is linearly computed by an adaptive weight.

Salient object detection on light fields. The idea of using light fields for saliency detection was proposed in Ref. [107]. A light field, captured using a specifically designed camera, e.g., Lytro, is essentially an array of images shot by a grid of cameras viewing the scene. Light field data offers two benefits for

salient object detection: 1) it allows synthesis of a stack of images focused at different depths, and 2) it provides an approximation of scene depth and occlusions.

With this additional information, Li et al. [107] first utilize the focus and objectness priors to robustly choose the background and select foreground candidates. Specifically, the layer with the estimated background likelihood score is used to estimate the background regions. Regions, coming from a mean-shift algorithm, with high foreground likelihood score are chosen as salient object candidates. Finally, the estimated background and foreground are utilized to compute a contrast-based saliency map on the all-focus image.

A new challenging benchmark dataset for light-field saliency analysis, known as HFUT-Lytro, was recently introduced in Ref. [140].

2.2 New testament: Deep learning based models

All methods reviewed so far use heuristics to detect salient objects. While hand-crafted features allow real-time detection performance, they suffer from several shortcomings that limit their ability to capture salient objects in challenging scenarios.

Convolutional neural networks (CNNs) [69], one of the most popular tools in machine learning, have been applied to many vision problems such as object recognition [141], semantic segmentation [70], and edge detection [142]. Recently, it has been shown that CNNs [44, 47] are also very effective when applied to salient object detection. Thanks to their multi-level and multi-scale features, CNNs are capable of accurately capturing the most salient regions without any prior knowledge (e.g., segment-level information). Furthermore, multi-level features allow CNNs to better locate the boundaries of the detected salient regions, even when shades or reflections exist. By exploiting the strong feature learning ability of CNNs, a series of algorithms has been proposed to learn saliency representations from large amounts of data. These CNN-based models continually improve upon the best results so far on almost all existing datasets, and are becoming the main stream solution. The rest of this subsection is dedicated to reviewing CNN-based models.

Basically, salient object detection models based on deep learning can be split into two main categories. The first category includes models that use multi-layer perceptrons (MLPs) for saliency detection. In these models, the input image is usually over-segmented into single- or multi-scale small regions. Then, a CNN is used to extract high-level features which are later fed to an MLP to determine the saliency value of each small region. Though high-level features are extracted from CNNs, unlike fully convolutional networks (FCNs), the spatial information from CNN features cannot be preserved because of the utilization of MLPs. To highlight the differences between these methods and FCN-based methods, we call them *classic convolutional network* based (CCN-based) methods. The second category includes models that are based on *fully convolutional networks* (FCN-based). The pioneering work of Long et al. [70] falls under this category and aims to solve the semantic segmentation problem. Since salient object detection is inherently a segmentation task, a number of researchers have adopted FCN-based architectures because of their ability to preserve spatial information.

Table 4 shows a list of CNN-based saliency models.

2.2.1 CCN-based models

One-dimensional convolution based methods. As an early attempt, He et al. [44] followed a region-based approach to learn superpixel-wise feature representations. Their approach dramatically reduces the computational cost compared to pixel-wise CNNs, while also taking global context into consideration. However, representing a superpixel with its mean color is not informative enough. Further, the spatial structure of the image is difficult to fully represent using 1D convolution and pooling operations, leading to cluttered predictions, especially when the input image is a complex scene.

Leveraging local and global context. Wang et al. consider both local and global information for better detection of salient regions [160]. To this end, two subnetworks are designed, one each for local estimation and global search. A deep neural network (DNN-L) is first used to learn local patch features to determine the saliency value of each pixel, followed by a refinement operation which captures high-level objectness. For global search, they train another deep neural network (DNN-G) to predict the saliency value of each salient region using a variety of global contrast features such as geometric information, etc.

Table 4 CNN-based salient object detection models and information used by them during training. Above: CCN-based models. Below: FCN-based models

#	Model	Pub	Year	#Training images	Training set	Pre-trained model	Fully conv
1	**SuperCNN** [44]	IJCV	2015	800	ECSSD	—	✗
2	**LEGS** [45]	CVPR	2015	3,340	MSRA-B+PASCALS	—	✗
3	**MC** [46]	CVPR	2015	8,000	MSRA10K	GoogLeNet [143]	✗
4	**MDF** [47]	CVPR	2015	2,500	MSRA-B	—	✗
5	**HARF** [48]	ICCV	2015	2,500	MSRA-B	—	✗
6	**ELD** [144]	CVPR	2016	nearly 9,000	MSRA10K	VGGNet	✗
7	**SSD-HS** [145]	ECCV	2016	2,500	MSRA-B	AlexNet	✗
8	**FRLC** [146]	ICIP	2016	4,000	DUT-OMRON	VGGNet	✗
9	**SCSD-HS** [147]	ICPR	2016	2,500	MSRA-B	AlexNet	✗
10	**DISC** [148]	TNNLS	2016	9,000	MSRA10K	—	✗
11	**LCNN** [149]	Neuro	2017	2,900	MSRA-B+PASCALS	AlexNet	✗
12	**DHSNET** [150]	CVPR	2016	6,000	MSRA10K	VGGNet	✓
13	**DCL** [151]	CVPR	2016	2,500	MSRA-B	VGGNet [152]	✓
14	**RACDNN** [153]	CVPR	2016	10,565	DUT+NJU2000+RGBD	VGG	✓
15	**SU** [154]	CVPR	2016	10,000	MSRA10K	VGGNet	✓
16	**CRPSD** [155]	ECCV	2016	10,000	MSRA10K	VGGNet	✓
17	**DSRCNN** [156]	MM	2016	10,000	MSRA10K	VGGNet	✓
18	**DS** [157]	TIP	2016	nearly 10,000	MSRA10K	VGGNet	✓
19	**IMC** [158]	WACV	2017	nearly 6,000	MSRA10K	ResNet	✓
20	**MSRNet** [159]	CVPR	2017	2,500	MSRA-B+HKU-IS	VGGNet	✓
21	**DSS** [49]	CVPR	2017	2,500	MSRA-B	VGGNet	✓

The top K candidate regions are utilized to compute the final saliency map using a weighted summation.

In Ref. [46], as in most classic salient object detection methods, both local context and global context are taken into account to construct a multi-context deep learning framework. The input image is first fed to the global-context branch to extract global contrast information. Meanwhile, each image patch, which is a superpixel-centered window, is fed to the local-context branch to capture local information. A binary classifier is finally used to determine the saliency value by minimizing a unified softmax loss between the prediction value and the ground truth label. A task-specific pre-training scheme is adopted to jointly optimize the designed multi-context model.

Lee et al. [144] exploit two subnetworks to encode low-level and high-level features separately. They first extract a number of features for each superpixel and feed them into a subnetwork composed of a stack of convolutional layers with 1×1 kernel size. Then, the standard VGGNet [152] is used to capture high-level features. Both low- and high-level features are flattened, concatenated, and finally fed into a two-layer MLP to judge the saliency of each query region.

Bounding box based methods. In Ref. [48], Zou and Komodakis propose a hierarchy-associated rich feature (HARF) extractor. A binary segmentation tree is first built to extract hierarchical image regions and to analyze the relationships between all pairs of regions. Two different methods are then used to compute two kinds of features (HARF_1 and HARF_2) for regions at the leaf-nodes of the binary segmentation tree. They leverage all the intermediate features extracted from the RCNN [161] to capture various characteristics of each image region. With these high-dimensional elementary features, both local regional contrast and border regional contrast for each elementary feature type are computed, to build a more compact representation. Finally, the AdaBoost algorithm is adopted to gradually assemble weak decision trees to construct a composite strong regressor.

Kim and Pavlovic [145] design a two-branch CNN architecture to obtain coarse- and fine-representations of coarse-level and fine-level patches, respectively. The selective search [162] method is utilized to generate a number of region candidates that are treated as input to the two-branch CNN. Feeding the concatenation of the feature representations of the two branches into the final fully connected layer allows a coarse continuous map to be predicted. To further refine the coarse prediction map, a hierarchical segmentation method is used to sharpen its boundaries and improve spatial consistency.

In Ref. [146], Wang et al. detect salient objects by employing the fast R-CNN [161] framework. The input image is first segmented into multi-scale regions

using both over-segmentation and edge-preserving methods. For each region, the external bounding box is used and the enclosed region is fed to the fast R-CNN. A small network composed of multiple fully connected layers is connected to the ROI pooling layer to determine the saliency value of each region. Finally, an edge-based propagation method is used to suppress background regions and make the resulting saliency map more uniform.

Kim and Pavlovic [147] train a CNN to predict the saliency shape of each image patch. The selective search method is first used to localize a stack of image patches, each of which is taken as input to the CNN. After predicting the shape of each patch, an intermediate mask M_I is computed by accumulating the product of the mask of the predicted shape class and the corresponding probability, and averaging all the region proposals. To further refine the coarse prediction map, shape class-based saliency detection with hierarchical segmentation (SCSD-HS) is used to incorporate more global information, which is often needed for saliency detection.

Li et al. [149] leverage both high-level features from CNNs and low-level features extracted using hand-crafted methods. To enhance the generalization and learning ability of CNNs, the original R-CNN is redesigned by adding local response normalization (LRN) to the first two layers. The selective search method is utilized [162] to generate a stack of square patches as the input to the network. Both high-level and low-level features are fed to an SVM with L_1 hinge-loss to help judge the saliency of each square region.

Models with multi-scale inputs. Li and Yu [47] utilize a pre-trained CNN as a feature extractor. Given an input image, they first decompose it into a series of non-overlapping regions and then feed them into a CNN with three different-scale inputs. Three subnetworks are then employed to capture advanced features at different scales. The features obtained from patches at three scales are concatenated and then fed into a small MLP with only two fully connected layers, using it as a regressor to output a distribution over binary saliency labels. To solve the problem of imperfect over-segmentation, a superpixel-based saliency refinement method is used.

Fiugre 4 illustrates a number of popular FCN-based architectures. Table 5 lists different types of information leveraged by these architectures.

Discussion. As can be seen, MLP-based works rely mostly on segment-level information (e.g., image patches) and classification networks. These image patches are normally resized to a fixed size and are then fed into a classification network which is used to determine the saliency of each patch. Some models use multi-scale inputs to extract features at several scales. However, such a learning framework cannot fully leverage high-level semantic information.

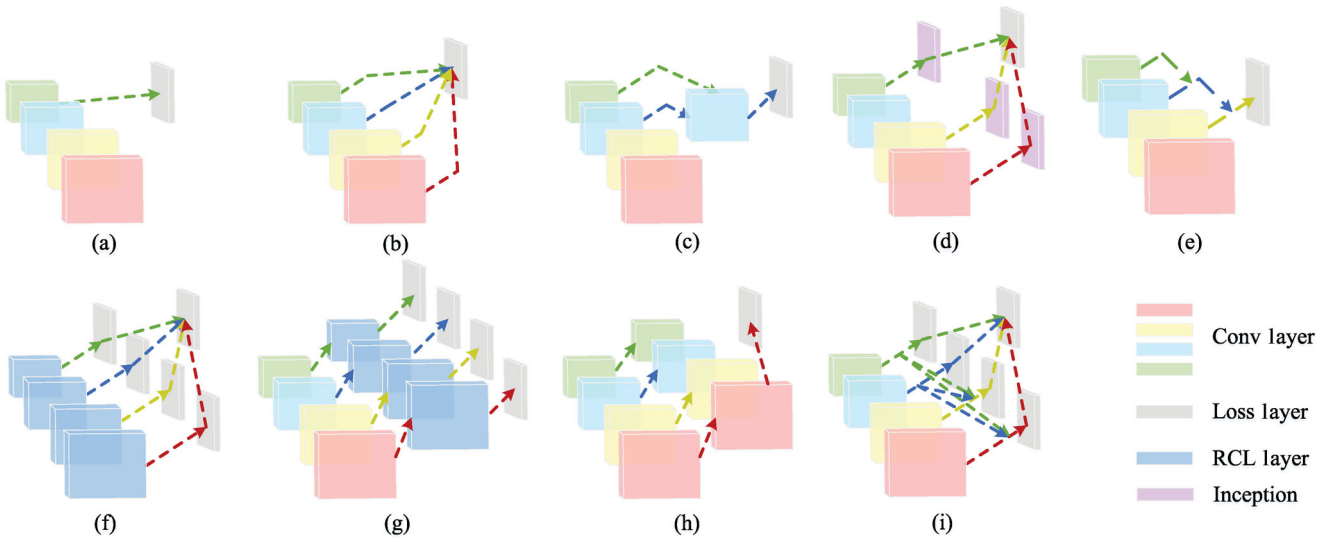


Fig. 4 Popular FCN-based architectures. Apart from the classical architecture (a), more and more advanced architectures have been developed recently. Some of them (b–e) exploit skip layers from different scales so as to learn multi-scale and multi-level features. Some (e, g–i) adopt an encoder–decoder structure to better fuse high-level features with low-level ones. Others (f, g, i) introduce side supervision as in Ref. [142] in order to capture more detailed multi-level information. See Table 5 for details of these architectures.

Table 5 Different types of information leveraged by existing FCN-based models. Abbreviations: SP: superpixel, SS: side supervision, RCL: recurrent convolutional layer, PCF: pure CNN feature, IL: instance-level, Arch: architecture

#	Model	SP	SS	RCL	PCF	IL	CRF	Arch.
1	**DCL** [151]	✓	✓	✗	✓	✗	✓	Fig. 4(b)
2	**CRPSD** [155]	✓	✗	✗	✗	✗	✗	Fig. 4(c)
3	**DSRCNN** [156]	✗	✓	✓	✓	✗	✗	Fig. 4(f)
4	**DHSNET** [150]	✗	✓	✓	✓	✗	✗	Fig. 4(g)
5	**RACDNN** [153]	✗	✗	✓	✓	✗	✗	Fig. 4(h)
6	**SU** [154]	✗	✓	✗	✓	✗	✓	Fig. 4(d)
7	**DS** [157]	✓	✗	✗	✗	✗	✗	Fig. 4(a)
8	**IMC** [158]	✓	✗	✗	✗	✗	✗	Fig. 4(a)
9	**MSRNet** [159]	✓	✗	✗	✓	✓	✓	Fig. 4(h)
10	**DSS** [49]	✗	✓	✗	✓	✗	✓	Fig. 4(i)

Further, spatial information cannot be propagated to the last fully connected layers, thus resulting in global information loss.

2.2.2 FCN-based models

Unlike CCN-based models that operate at the patch level, fully convolutional networks (FCNs) [70] consider pixel-level operations to overcome problems caused by fully connected layers such as blurring and inaccurate predictions near the boundaries of salient objects. Due to the desirable properties of FCNs, a great number of FCN-based salient object detection models have been introduced recently.

Li and Yu [151] design a CNN with two complementary branches: a *pixel-level fully convolutional stream* (FCS) and a *segment-wise spatial pooling stream* (SPS). The FCS introduces a series of skip layers after the last convolutional layer of each stage; the skip layers are fused together as the output of the FCS. Note that a stage of the CNN is composed of all layers with the same resolution. The SPS leverages segment-level information for spatial pooling. Finally, the outputs of FCS and SPS are fused, followed by a balanced sigmoid cross entropy loss layer as used in Ref. [142].

Liu and Han [150] propose two subnetworks to produce a prediction map working in a coarse-to-fine and global-to-local manner. The first subnetwork can be considered as an encoder whose goal is to generate a coarse global prediction. Then, a refinement subnetwork composed of a series of recurrent convolution layers is used to refine the coarse prediction map from coarse scales to fine scales.

In Ref. [155], Tang and Wu consider both region-level saliency estimation and pixel-level saliency prediction. For pixel-level prediction, two side paths are connected to the last two stages of the VGGNet and then concatenated to learn multi-scale features. For region-level estimation, each given image is first over-segmented into multiple superpixels and then the Clarifai model [163] is used to predict the saliency of each superpixel. The original image and the two prediction maps are taken as the inputs to a small CNN to generate a more convincing saliency map as the final output.

Tang et al. [156] take the deeply supervised net [164] and adopt a similar architecture as in the holistically-nested edge detector [142]. Unlike HED, they replace the original convolutional layers in VGGNet with recurrent convolutional layers to learn local, global, and contextual information.

In Ref. [153], Kuen et al. propose a two-stage CNN by utilizing spatial transformer and recurrent network units. A convolutional–deconvolutional network is first used to produce an initial coarse saliency map. The spatial transformer network [165] is applied to extract multiple sub-regions from the original images, followed by a series of recurrent network units to progressively refine the predictions of these sub-regions.

Kruthiventi et al. [154] consider both fixation prediction and salient object detection in a unified network. To capture multi-scale semantic information, four inception modules [143] are introduced which are connected to the output of the 2nd, 4th, 5th, and 6th stages, respectively. These four side paths are concatenated and passed through a small network composed of two convolutional layers to reduce the aliasing effect of upsampling. Finally, the sigmoid cross entropy loss is used to optimize the model.

Li et al. [157] consider joint semantic segmentation and salient object detection. As in the FCN work [70], the two original fully connected layers in VGGNet [152] are replaced by convolutional layers. To overcome the fuzzy object boundaries caused by the down-sampling operations of CNNs, they make use of the SLIC [166] superpixels to model the topological relationships between superpixels in both spatial and feature dimensions. Finally, graph Laplacian regularized nonlinear regression is used to change the combination of the predictions from CNNs and the superpixel graph from the coarse level to the fine level.

Zhang et al. [158] detect salient objects using saliency cues extracted by CNNs and a multi-level fusion mechanism. The Deeplab [167] architecture is first used to capture high-level features. To address the problem of large strides in Deeplab, a multi-scale binary pixel labeling method is adopted to improve spatial coherence, as in Ref. [47].

The MSRNet [159] by Li et al. performs both salient object detection and instance-level salient object segmentation. A multi-scale CNN is used to simultaneously detect salient regions and contours. For each scale, features from upper layers are merged with features from lower layers to gradually refine the results. To generate a contour map, the MCG [168] approach is used to extract a small number of candidate bounding boxes and well-segmented regions that are used to help perform salient object instance segmentation. Finally, a fully connected CRF model [169] is employed to refine the spatial coherence.

Hou et al. [49] design a top–down model based on the HED architecture [142]. Instead of connecting independent side paths to the last convolutional layer of each stage, a series of short connections are introduced to build a strong relationship between each pair of side paths. As a result, features from upper layers with strongly semantic information are propagated to lower layers, helping them accurately locate exact positions of salient objects. In the meantime, rich detailed information from lower layers allow irregular prediction maps from deeper layers to be refined. A special fusion mechanism is exploited to better combine the saliency maps predicted by different side paths.

Discussion. The foregoing approaches are all based on fully convolutional networks, which enable point-to-point learning and end-to-end training strategies. Compared to CCN-based models, these methods make better use of the convolution operation and substantially decrease the time cost. More importantly, recent FCN-based approaches [49, 159] that utilize CNN features greatly outperform those methods with segment-level information.

To sum up, the three following advantages are obtained in utilizing FCN-based models for saliency detection:

1. **Local versus global**. As mentioned in Section 2.2.1, earlier CNN-based models incorporate both local and global contextual information explicitly (embedded in separate networks [45–47]) or implicitly (using an end-to-end framework). This indeed agrees with the design principles behind many hand-crafted cues reviewed in previous sections. However, FCN-based methods are capable of learning both local and global information internally. Lower layers tend to encode more detailed information such as edge and fine components, while deeper layers favor global and semantically meaningful information. Such properties enable FCN-based networks to drastically outperform classic methods.
2. **Pre-training and fine-tuning**. The effectiveness of fine-tuning a pre-trained network has been demonstrated in many different applications. The network is typically pre-trained on the ImageNet dataset [170] for image classification. The learned knowledge can be applied to several different target tasks (e.g., object detection [161], object localization [171]) through simple fine-tuning. A similar strategy has been adopted for salient object detection [46, 151] and has resulted in superior performance compared to training from scratch. The learned features, more importantly, are able to capture high-level semantic knowledge about object categories, as the employed networks are pre-trained for scene and object classification tasks.
3. **Versatile architectures**. A CNN architecture is formed by a stack of distinct layers that transform the input images into an output map through a differentiable function. The diversity of FCNs allows designers to design different structures that are appropriate for them.

Despite great success, FCN-based models still fail in several cases. Typical examples include scenes with transparent objects, low contrast between foreground and background, and complex backgrounds, as shown in Ref. [49]. This calls for development of more powerful architectures in future.

Figure 5 provides a visual comparison of maps generated by classic and CNN-based models.

3 Applications of salient object detection

The value of salient object detection models lies in their application to many areas of computer vision, graphics,

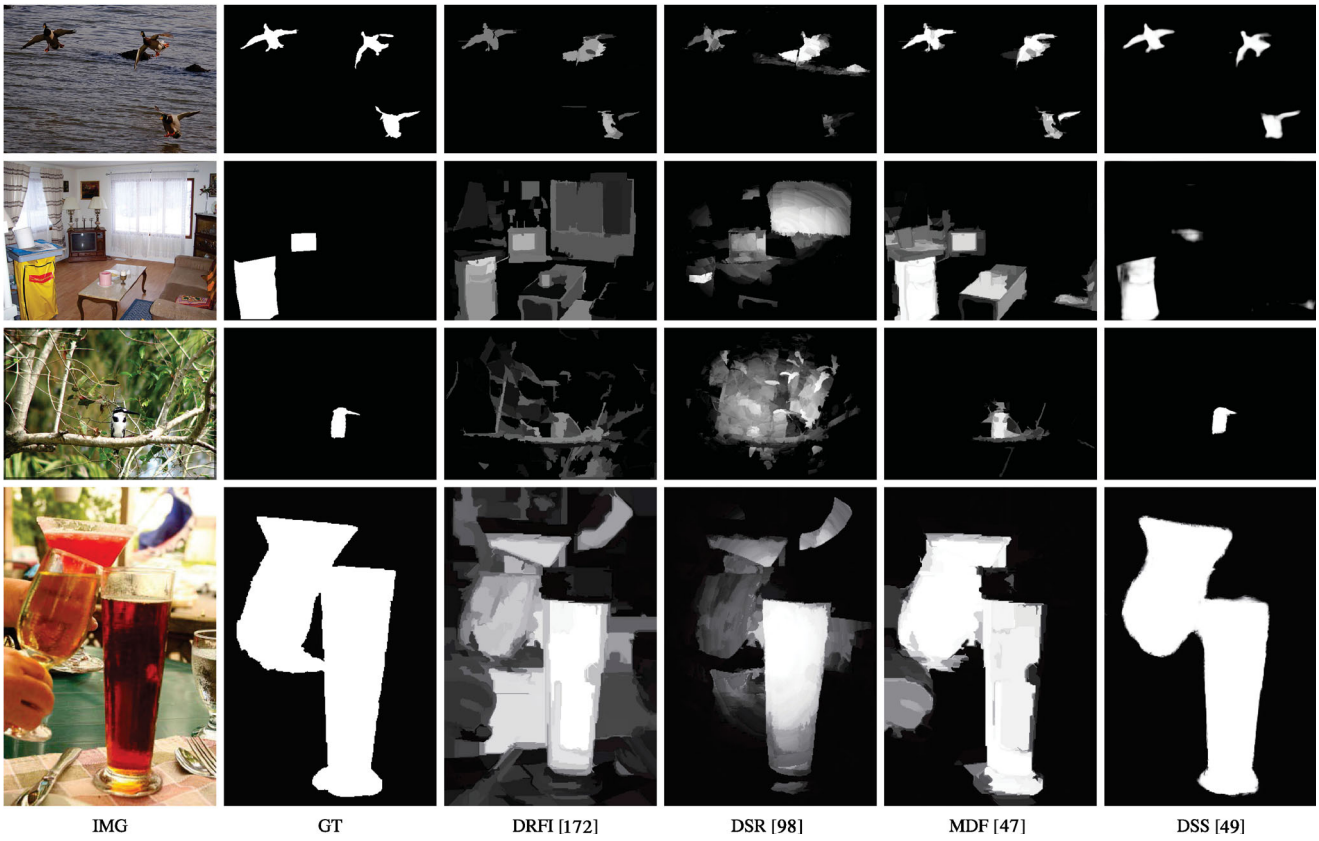


Fig. 5 Visual comparisons of two best classic methods (DRFI and DSR), according to Ref. [132], and two leading CNN-based methods (MDF and DSS).

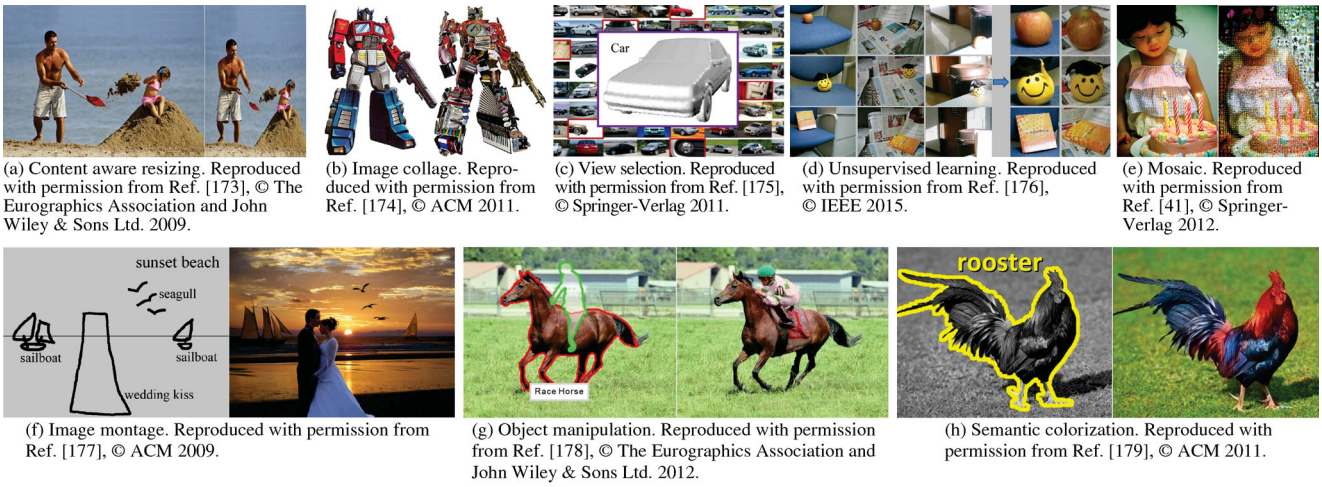
and robotics. Salient object detection models have been utilized for several applications such as object detection and recognition [180–186], image and video compression [187, 188], video summarization [189–191], photo collage/media re-targeting/cropping/thumbnailing [174, 192, 193], image quality assessment [194–196], image segmentation [197–200], content-based image retrieval and image collection browsing [177, 201–203], image editing and manipulation [41, 175, 178, 179], visual tracking [204–210], object discovery [211, 212], and human-robot interaction [213, 214]. Figure 6 shows example applications.

4 Datasets and evaluation measures

4.1 Salient object detection datasets

As more and more models have been proposed in the literature, more datasets have been introduced to further challenge saliency detection models. Early attempts aim to collect images with salient objects being annotated with bounding boxes (e.g., *MSRA-A* and *MSRA-B* [25]), while later efforts annotate such salient objects with pixel-wise binary masks (e.g., *ASD* [37] and *DUT-OMRON* [97]). Typically, images, which can be annotated with accurate masks, contain few objects (usually one) and simple background regions. On the contrary, recent attempts have been made to collect datasets with multiple objects in complex scenes with cluttered backgrounds (e.g., Refs. [22, 23, 26]). As already noted, a more sophisticated mechanism is required to determine the most salient object when several candidate objects are present in the same scene. For example, Borji [23] and Li et al. [22] use the peak of the human fixation map to determine which object is the most salient (i.e., the one that humans look at the most; see Section 1.2).

A list of 22 salient object datasets including 20 image datasets and 2 video datasets is provided in Table 6. Notice that all images or video frames in these datasets are annotated with binary masks or rectangles. Subjects are often asked to label a single salient object in an image (e.g., Ref. [25]) or to annotate the most salient among several candidate



(a) Content aware resizing. Reproduced with permission from Ref. [173], © The Eurographics Association and John Wiley & Sons Ltd. 2009.

(b) Image collage. Reproduced with permission from Ref. [174], © ACM 2011.

(c) View selection. Reproduced with permission from Ref. [175], © Springer-Verlag 2011.

(d) Unsupervised learning. Reproduced with permission from Ref. [176], © IEEE 2015.

(e) Mosaic. Reproduced with permission from Ref. [41], © Springer-Verlag 2012.

(f) Image montage. Reproduced with permission from Ref. [177], © ACM 2009.

(g) Object manipulation. Reproduced with permission from Ref. [178], © The Eurographics Association and John Wiley & Sons Ltd. 2012.

(h) Semantic colorization. Reproduced with permission from Ref. [179], © ACM 2011.

Fig. 6 Sample applications of salient object detection.

Table 6 Overview of popular salient object datasets. Above: image datasets, below: video datasets. Obj: objects per image, Ann: Annotation, Sbj: Subjects/Annotators, Eye: Eye tracking subjects, I/V: Image/Video

Dataset	Year	Imgs	Obj	Ann	Resolution	Sbj	Eye	I/V
MSRA-A [25, 215]	2007	20k	∼1	BB	400 × 300	3	—	I
MSRA-B [25, 215]	2007	5k	∼1	BB	400 × 300	9	—	I
SED1 [132, 216]	2007	100	1	PW	∼300 × 225	3	—	I
SED2 [132, 216]	2007	100	2	PW	∼300 × 225	3	—	I
ASD [25, 37]	2009	1000	∼1	PW	400 × 300	1	—	I
SOD [60, 217]	2010	300	∼3	PW	481 × 321	7	—	I
iCoSeg [125]	2010	643	∼1	PW	∼500 × 400	1	—	I
MSRA5K [25, 93]	2011	5k	∼1	PW	400 × 300	1	—	I
Infrared [218, 219]	2011	900	∼5	PW	1024 × 768	2	15	I
ImgSal [205]	2013	235	∼2	PW	640 × 480	19	50	I
CSSD [42]	2013	200	∼1	PW	∼400 × 300	1	—	I
ECSSD [42, 220]	2013	1000	∼1	PW	∼400 × 300	1	—	I
MSRA10K [25, 221]	2013	10k	∼1	PW	400 × 300	1	—	I
THUR15K [25, 221]	2013	15k	∼1	PW	400 × 300	1	—	I
DUT-OMRON [97]	2013	5,172	∼5	BB	400 × 400	5	5	I
Bruce-A [26, 54]	2013	120	∼4	PW	681 × 511	70	20	I
Judd-A [23, 222]	2014	900	∼5	PW	1024 × 768	2	15	I
PASCAL-S [22]	2014	850	∼5	PW	Variable	12	8	I
UCSB [223]	2014	700	∼5	PW	405 × 405	100	8	I
OSIE [224]	2014	700	∼5	PW	800 × 600	1	15	I
RSD [225]	2009	62,356	Var.	BB	Variable	23	—	V
STC [226]	2011	4,870	∼1	BB	Variable	1	—	V

objects (e.g., Ref. [26]). Some image datasets also provide for each image the fixation data collected during a free-viewing task.

4.2 Evaluation measures

Five universally-agreed, standard, and easy-to-compute measures for evaluating salient object detection models are described next. For simplicity, we use S to represent the predicted saliency map normalized to $[0, 255]$ and G to be the ground-truth binary mask of salient objects. For a binary mask, we use $|\cdot|$ to represent the number of non-zero entries in the mask.

4.2.1 Precision–recall (PR)

A saliency map S is first converted to a binary mask M and then *Precision* and *Recall* are computed by comparing M to the ground-truth G:

$$\text{Precision} = \frac{|M \cap G|}{|M|}, \quad \text{Recall} = \frac{|M \cap G|}{|G|} \tag{11}$$

Binarization of S is the key step in the evaluation. There are three popular ways to perform binarization. In the first solution, Achanta et al. [37] propose image-dependent adaptive threshold for binarizing S, computed as twice as the mean saliency of S:

$$T_{\text{a}} = \frac{2}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} S(x, y) \tag{12}$$

where W and H are the width and the height of the saliency map S, respectively.

The second way to binarize S is to use a threshold that varies from 0 to 255. For each threshold, a pair of (precision, recall) scores are computed and used to plot a precision–recall (PR) curve.

The third way to perform binarization is to use a GrabCut-like algorithm (e.g., as in Ref. [84]). Here, the PR curve is first computed and the threshold that leads to 95% recall is selected. With this threshold, an initial binary mask is generated, which is then used to initialize iterative GrabCut segmentation [138] to gradually refine the binary mask.

4.2.2 F-measure

Often, neither precision nor recall can fully evaluate the quality of a saliency map. Instead, the F-measure is used, defined as the weighted harmonic mean of precision and recall with a non-negative weight β^2:

$$F_\beta = \frac{(1+\beta^2)\ \text{Precision} \times \text{Recall}}{\beta^2\ \text{Precision} + \text{Recall}} \tag{13}$$

In many salient object detection works (e.g., Ref. [37]), β^2 is set to 0.3 to give greater weight to precision: recall rate is not as important as precision (see also Ref. [55]). For instance, 100% recall can be easily achieved by setting the whole map to be foreground.

4.2.3 Receiver operating characteristics (ROC) curve

In the above, false positive rate (FPR) and true positive rate (TPR) can be computed when binarizing the saliency map with a set of fixed thresholds:

$$\text{TPR} = \frac{|M \cap G|}{|G|}, \quad \text{FPR} = \frac{|M \cap G|}{|M \cap G| + |\bar{M} \cap \bar{G}|} \tag{14}$$

where $\bar{M}$ and $\bar{G}$ denote the complement of the binary mask M and ground-truth G, respectively. The ROC curve is the plot of TPR versus FPR for all possible thresholds.

4.2.4 Area under ROC curve (AUC)

While the ROC is a 2D representation of a model's performance, the AUC distils this information into a single number. As the name implies, it is calculated as the area under the ROC curve. A perfect model will score an AUC of 1, while random guessing will score an AUC of around 0.5.

4.2.5 Mean absolute error (MAE)

The overlap-based evaluation measures introduced above do not consider true negative saliency assignments, i.e., the pixels correctly marked as non-salient. They favor methods that successfully assign high saliency to salient pixels but fail to detect non-salient regions. Moreover, for some applications [227], the quality of the weighted continuous saliency maps may be of higher interest than the binary masks. For a more comprehensive comparison, it is recommended to evaluate the mean absolute error (MAE) between the continuous saliency map S and the binary ground-truth G, both normalized to the range [0, 1]. The MAE score is defined as

$$\text{MAE} = \frac{1}{WH} \sum_{x=1}^{W} \sum_{y=1}^{H} \parallel S(x,y) - G(x,y) \parallel \tag{15}$$

Please refer to Ref. [228] for more details on datasets and scores in the filed of salient object detection. Code for evaluation measures is available at `http://mmcheng.net/salobjbenchmark`.

5 Discussion

5.1 Design choices

In the past two decades, hundreds of classic and deep learning based methods have been proposed for detecting and segmenting salient objects in scenes, and a large number of design choices have been explored. Although great success has been achieved recently, there is still large room for improvement. Our detailed method summarization (see Table 1 and Table 2) sends some clear messages about the commonly used design choices, and these are valuable for the design of future algorithms, as we now discuss.

5.1.1 Heuristics versus learning from data

Early methods were mainly based on heuristic cues (local or global) to detect salient objects [27, 37, 84, 97]. Recently, saliency models based on learning algorithms have shown to be very effective (see Table 1 and Table 2). Among these models, deep learning based methods greatly outperform conventional heuristic methods because of their ability to learn large amounts of extrinsic cues from large datasets. Data-driven approaches for salient object detection seem to have surprisingly good generalization ability. An emerging question, however, is whether the data-driven ideas for salient object detection conflict with the ease of use of these models. Most learning based approaches are only trained on a small subset of the **MSRA5K** dataset, and still consistently outperform other methods on all other datasets which have

considerable differences. This suggests that it is worth further exploring data-driven salient object detection without losing the advantages of simplicity and ease-of-use, in particular from an application point of view.

5.1.2 Hand-crafted versus CNN-based features

The first generation of learning-based methods were based on many hand-crafted features. An obvious drawback of these methods is their generalizability, especially when applied to complex cluttered scenes. In addition, these methods mainly rely on over-segmentation algorithms, such as SLIC [166], yielding incomplete salient objects having high contrast components. CNN-based models solve these problems, to some degree, even when complex scenes are considered. Because of their ability to learn multi-level features, it is easy for CNNs to accurately locate salient objects. Low-level features such as edges enable sharpening boundaries of salient objects while high-level features allow incorporating semantic information to identify salient objects.

5.1.3 Recent advances in CNN-based saliency detection

Various CNN-based architectures have been proposed recently. Among these approaches, there are several promising choices that can be further explored in future. The first one regards models with deep supervision. As shown in Ref. [49], deeply supervised networks strengthen the power of features in different layers. The second choice is the encoder–decoder architecture, which has been adopted in many segmentation-related tasks. Such approaches gradually back-propagate high-level features to lower layers, allowing effective fusion of multi-level features. Another choice is to exploit stronger baseline models, such as using very deep ResNets [229] instead of VGGNet [152].

5.2 Dataset bias

Datasets have been important in the rapid progress in saliency detection. On one hand, they supply large scale training data and enable performance comparisons of competing algorithms. On the other hand, each dataset is a unique sampling of an unlimited application domain, and contains a certain degree of bias.

To date, there seems to be a unanimous agreement on the presence of bias (i.e., skew) in underlying structures of datasets. Consequently, some studies have addressed the effect of bias in image datasets. For instance, Torralba and Efros identify three biases in computer vision datasets, namely: *selection bias, capture bias*, and *negative set bias* [230]. Selection bias is caused by preference for a particular kind of image during data gathering. It results in qualitatively similar images in a dataset. This is witnessed by the strong color contrast (see Refs. [22, 84]) in most frequently used salient object benchmark datasets [37]. Thus, two practices in dataset construction are to be preferred: i) *having independent image selection and annotation processes* [22], and ii) *detecting the most salient object first and then segmenting it.* Negative set bias is the consequence of a lack of a rich and unbiased negative set, i.e., one should avoid concentrating on a particular image of interest and datasets should represent the whole world. Negative set bias may affect the ground-truth by incorporating the annotator's personal preferences for some object types. Thus, including a variety of images is encouraged when constructing a good dataset. Capture bias conveys the effect of image composition on the dataset. The most popular kind of such a bias is the tendency to compose images with important objects in the central region of the image, i.e., center bias. The existence of bias in a dataset makes quantitative comparisons very challenging and sometimes even misleading. For instance, a trivial saliency model which consists of a Gaussian blob at the image center often scores higher than many fixation prediction models [63, 231, 232].

5.3 Future directions

Several promising research directions for constructing more effective models and benchmarks are discussed here.

5.3.1 Beyond single images

Most benchmarks and saliency models discussed in this study deal with single images. Unfortunately, salient object detection on multiple input images, e.g., salient object detection on video sequences, co-salient object detection, and salient object detection over depth and light field images, are less explored. One reason behind this is the limited availability of benchmark datasets for these problems. For example, as mentioned in Section 4, there are only two publicly available benchmark datasets for video saliency (mostly comprising cartoons and news). For

these videos, only bounding boxes are provided for the key frames to roughly localize salient objects. Multi-modal data is becoming increasingly more accessible and affordable. Integrating additional cues such as spatio-temporal consistency and depth will be beneficial for efficient salient object detection.

5.3.2 Instance-level salient object detection

Existing saliency models are object-agnostic (i.e., they do not split salient regions into objects). However, humans possess the capability to detect salient objects at instance level. Instance-level saliency can be useful in several applications, such as image editing and video compression.

Two possible approaches for instance-level saliency detection are as follows. The first uses an object detection or object proposal method, e.g., Fast-RCNN [161], to extract a stack of object bounding box candidates and then segment salient objects within them. The second approach, initially proposed in Ref. [159], is to leverage edge information to distinguish different salient objects.

5.3.3 Versatile network architectures

With the deeper understanding of researchers on CNNs, more and more interesting network architectures have been developed. Using advanced baseline models and network architectures [151] can substantially improve the performance. On one hand, deeper networks help better capture salient objects because of their ability to extract high-level semantic information. On the other hand, apart from high-level information, low-level features [49, 159] should also be considered to build high resolution saliency maps.

5.3.4 Unanswered questions

Some remaining questions include: how many (salient) objects are necessary to represent a scene? Does map smoothing affect the scores and model ranking? How is salient object detection different from other fields? What is the best way to tackle center bias in model evaluation? What is the remaining gap between models and humans? A collaborative engagement with other related fields such as saliency for fixation prediction, scene labeling and categorization, semantic segmentation, object detection, and object recognition can help answer these questions, situate the field better, and identify future directions.

6 Summary and conclusions

In this paper, we have exhaustively reviewed the salient object detection literature with respect to closely related areas. Detecting and segmenting salient objects is very useful. Objects in images automatically capture more attention than background items, such as grass, trees, and sky. Therefore, if we can detect salient or important objects first, we can perform detailed reasoning and scene understanding in the next stage. Compared to traditional special-purpose object detectors, saliency models are general, typically fast, and do not need heavy annotation. These properties allow processing of a large number of images at low cost.

Exploring connections between salient object detection and fixation prediction models can help enhance performance for both types of models. In this regard, datasets that offer both salient object judgements of humans and eye movements are highly desirable. Conducting behavioral studies to understand how humans perceive and prioritize objects in scenes and how this concept is related to language, scene description and captioning, visual question answering, attributes, etc., can offer invaluable insights. Further, it is critical to focus more on evaluating and comparing salient object models to gauge future progress. Tackling dataset biases such as center bias and selection bias and moving towards more challenging images is important.

Although salient object detection and segmentation methods have made great strides in recent years, a very robust salient object detection algorithm that can generate high quality results for nearly all images is still lacking. Even for humans, what is the most salient object in the image, is sometimes a quite ambiguous question. To this end, a general suggestion:

Don't ask what segments can do for you, ask what you can do for the segments[①].

— Jitendra Malik

is particularly important when attempting to build robust algorithms. For instance, when dealing with noisy Internet images, although salient object detection and segmentation methods do not guarantee robust performance on individual images,

① http://www.cs.berkeley.edu/$\sim$malik/student-tree-2010.pdf

their efficiency and simplicity make it possible to automatically process a large number of images. This allows the filtering of images for the purposes of reliability and accuracy, running applications robustly [84, 174, 175, 177, 179, 233], and unsupervised learning [176].

References

[1] Cheng, M.; Mitra, N. J.; Huang, X.; Torr, P. H. S.; Hu, S. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 37, No. 3, 569–582, 2015.

[2] Bylinskii, Z.; Judd, T.; Borji, A.; Itti, L.; Durand, F.; Oliva, A.; Torralba, A. MIT saliency benchmark. 2015. Available at `http://saliency.mit.edu/results_mit300.html`.

[3] Bylinskii, Z.; Recasens, A.; Borji, A.; Oliva, A.; Torralba, A.; Durand, F. Where should saliency models look next? In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9909*. Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 809–824, 2016.

[4] Spain, M.; Perona, P. Measuring and predicting object importance. *International Journal of Computer Vision* Vol. 91, No. 1, 59–76, 2011.

[5] Berg, A. C.; Berg, T. L.; Daume, H.; Dodge, J.; Goyal, A.; Han, X.; Mensch, A.; Mitchell, M.; Sood, A.; Stratos, K. et al. Understanding and predicting importance in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3562–3569, 2012.

[6] M't Hart, B. M.; Schmidt, H. C. E. F.; Roth, C.; Einhäuser, W. Fixations on objects in natural scenes: Dissociating importance from salience. *Frontiers in Psychology* Vol. 4, 455, 2013.

[7] Isola, P.; Xiao, J.; Torralba, A.; Oliva, A. What makes an image memorable? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 145–152, 2011.

[8] Rosenholtz, R.; Li, Y. Z.; Nakano, L. Measuring visual clutter. *Journal of Vision* Vol. 7, No. 2, 17, 2007.

[9] Katti, H.; Bin, K. Y.; Chua, T. S.; Kankanhalli, M. Preattentive discrimination of interestingness in images. In: Proceedings of the IEEE International Conference on Multimedia and Expo, 1433–1436, 2008.

[10] Gygli, M.; Grabner, H.; Riemenschneider, H.; Nater, F.; Van Gool, L. The interestingness of images. In: Proceedings of the IEEE International Conference on Computer Vision, 1633–1640, 2013.

[11] Dhar, S.; Ordonez, V.; Berg, T. L. High level describable attributes for predicting aesthetics and interestingness. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1657–1664, 2011.

[12] Jiang, Y.-G.; Wang, Y.; Feng, R.; Xue, X.; Zheng, Y.; Yang, H. Understanding and predicting interestingness of videos. In: Proceedings of the 27th AAAI Conference on Artificial Intelligence, 2013.

[13] Itti, L.; Baldi, P. Bayesian surprise attracts human attention. In: Proceedings of the 18th International Conference on Neural Information Processing Systems, 547–554, 2005.

[14] Wang, Z.; Bovik, A. C.; Sheikh, H. R.; Simoncelli, E. P. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing* Vol. 13, No. 4, 600–612, 2004.

[15] Wang, Z.; Bovik, A. C.; Lu, L. Why is image quality assessment so difficult? In: Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing, IV-3313–IV-3316, 2002.

[16] Zhang, W.; Borji, A.; Wang, Z.; Le Callet, P.; Liu, H. T. The application of visual saliency models in objective image quality assessment: A statistical evaluation. *IEEE Transactions on Neural Networks and Learning Systems* Vol. 27, No. 6, 1266–1278, 2016.

[17] Vogel, J.; Schiele, B. A semantic typicality measure for natural scene categorization. In: *Pattern Recognition. Lecture Notes in Computer Science, Vol. 3175*. Rasmussen, C. E.; Bülthoff, H. H.; Schölkopf, B.; Giese, M. A. Eds. Springer Berlin Heidelberg, 195–203, 2004.

[18] Ehinger, K. A.; Xiao, J.; Torralba, A.; Oliva, A. Estimating scene typicality from human ratings and image features. In: Proceedings of the Annual Cognitive Science Conference, 2011.

[19] Farhadi, A.; Endres, I.; Hoiem, D.; Forsyth, D. Describing objects by their attributes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1778–1785, 2009.

[20] Liu, H. Y.; Jiang, S. Q.; Huang, Q. M.; Xu, C. S.; Gao, W. Region-based visual attention analysis with its application in image browsing on small displays. In: Proceedings of the 15th ACM International Conference on Multimedia, 305–308, 2007.

[21] Mishra, A. K.; Aloimonos, Y.; Cheong, L. F.; Kassim, A. Active visual segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 34, No. 4, 639–653, 2012.

[22] Li, Y.; Hou, X.; Koch, C.; Rehg, J. M.; Yuille, A. L. The secrets of salient object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 280–287, 2014.

[23] Borji, A. What is a salient object? A dataset and a baseline model for salient object detection. *IEEE Transactions on Image Processing* Vol. 24, No. 2, 742–756, 2015.

[24] Itti, L.; Koch, C.; Niebur, E. A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 20, No. 11, 1254–1259, 1998.

[25] Liu, T.; Sun, J.; Zheng, N.; Tang, X.; Shum, H.-Y. Learning to detect a salient object. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–8, 2007.

[26] Borji, A.; Sihite, D. N.; Itti, L. What stands out in a scene? A study of human explicit saliency judgment. *Vision Research* Vol. 91, 62–77, 2013.

[27] Perazzi, F.; Krahenbuhl, P.; Pritch, Y.; Hornung, A. Saliency filters: Contrast based filtering for salient region detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 733–740, 2012.

[28] Comaniciu, D.; Meer, P. Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 24, No. 5, 603–619, 2002.

[29] Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 34, No. 11, 2274–2282, 2012.

[30] Cheng, M.-M.; Zhang, Z.; Lin, W.-Y.; Torr, P. H. S. BING: Binarized normed gradients for objectness estimation at 300fps. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3286–3293, 2014.

[31] Borji, A.; Itti, L. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 35, No. 1, 185–207, 2013.

[32] Borji, A.; Tavakoli, H. R.; Sihite, D. N.; Itti, L. Analysis of scores, datasets, and models in visual saliency prediction. In: Proceedings of the IEEE International Conference on Computer Vision, 921–928, 2013.

[33] Hosang, J.; Benenson, R.; Dollár, P.; Schiele, B. What makes for effective detection proposals? *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 38, No. 4, 814–830, 2016.

[34] Alexe, B.; Deselaers T.; Ferrari, V. What is an object? In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 73–80, 2010.

[35] Siva, P.; Russell, C.; Xiang, T.; Agapito, L. Looking beyond the image: Unsupervised learning for object saliency and detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3238–3245, 2013.

[36] Cheng, H. D.; Jiang, X. H.; Sun, Y.; Wang, J. L. Color image segmentation: Advances and prospects. *Pattern Recognition* Vol. 34, No. 12, 2259–2281, 2001.

[37] Achanta, R.; Hemami, S.; Estrada, F.; Süsstrunk, S. Frequency-tuned salient region detection. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 1597–1604, 2009.

[38] Cheng, M.-M.; Zhang, G.-X.; Mitra, N. J.; Huang, X.; Hu, S.-M. Global contrast based salient region detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 409–416, 2011.

[39] Goferman, S.; Zelnik-Manor, L.; Tal, A. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 34, No. 10, 1915–1926, 2012.

[40] Jiang, H. Z.; Wang, J. D.; Yuan, Z. J.; Wu, Y.; Zheng, N. N.; Li, S. P. Salient object detection: A discriminative regional feature integration approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2083–2090, 2013.

[41] Margolin, R.; Zelnik-Manor, L.; Tal, A. Saliency for image manipulation. *The Visual Computer* Vol. 29, No. 5, 381–392, 2013.

[42] Yan, Q.; Xu, L.; Shi, J.; Jia, J. Hierarchical saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1155–1162, 2013.

[43] Yang, C.; Zhang, L. H.; Lu, H. C. Graph-regularized saliency detection with convex-hull-based center prior. *IEEE Signal Processing Letters* Vol. 20, No. 7, 637–640, 2013.

[44] He, S.; Lau, R. W. H.; Liu, W.; Huang, Z.; Yang, Q. SuperCNN: A superpixelwise convolutional neural network for salient object detection. *International Journal of Computer Vision* Vol. 115, No. 3, 330–344, 2015.

[45] Wang, L.; Lu, H.; Ruan, X.; Yang, M.-H. Deep networks for saliency detection via local estimation and global search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3183–3192, 2015.

[46] Zhao, R.; Ouyang, W.; Li, H.; Wang, X. Saliency detection by multi-context deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1265–1274, 2015.

[47] Li, G.; Yu, Y. Visual saliency based on multiscale deep features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5455–5463, 2015.

[48] Zou, W.; Komodakis, N. HARF: Hierarchy-associated rich features for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision, 406–414, 2015.

[49] Hou, Q.; Cheng, M.-M.; Hu, X.; Borji, A.; Tu, Z.; Torr, P. H. S. Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3203–3212, 2017.

[50] Treisman, A. M.; Gelade, G. A feature-integration theory of attention. *Cognitive Psychology* Vol. 12, No. 1, 97–136, 1980.

[51] Wolfe, J. M.; Cave, K. R.; Franzel, S. L. Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance* Vol. 15, No. 3, 419–433, 1989.

[52] Koch, C.; Ullman, S. Shifts in selective visual attention: Towards the underlying neural circuitry. In: *Matters of Intelligence. Synthese Library (Studies in Epistemology, Logic, Methodology, and Philosophy of Science), Vol. 188.* Vaina, L. M. Ed. Springer Dordrecht, 115–141, 1987.

[53] Parkhurst, D.; Law, K.; Niebur, E. Modeling the role of salience in the allocation of overt visual attention. *Vision Research* Vol. 42, No. 1, 107–123, 2002.

[54] Bruce, N. D. B.; Tsotsos, J. K. Saliency based on information maximization. In: Proceedings of the 18th International Conference on Neural Information Processing Systems, 155–162, 2005.

[55] Liu, T.; Yuan, Z. J.; Sun, J.; Wang, J. D.; Zheng, N. N.; Tang, X. O.; Shum, H.-Y. Learning to detect a salient object. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 33, No. 2, 353–367, 2011.

[56] Achanta, R.; Estrada, F.; Wils, P.; Süsstrunk, S. Salient region detection and segmentation. In: *Computer Vision Systems. Lecture Notes in Computer Science, Vol. 5008.* Gasteratos, A.; Vincze, M.; Tsotsos, J. K. Eds. Springer Berlin Heidelberg, 66–75, 2008.

[57] Ma, Y.-F.; Zhang, H.-J. Contrast-based image attention analysis by using fuzzy growing. In: Proceedings of the 11th ACM International Conference on Multimedia, 374–381, 2003.

[58] Liu, F.; Gleicher, M. Region enhanced scale-invariant saliency detection. In: Proceedings of the IEEE International Conference on Multimedia and Expo, 1477–1480, 2006.

[59] Walther, D.; Koch, C. Modeling attention to salient proto-objects. *Neural Networks* Vol. 19, No. 9, 1395–1407, 2006.

[60] Arbeláez, P.; Maire, M.; Fowlkes, C.; Malik, J. Contour detection and hierarchical image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 33, No. 5, 898–916, 2011.

[61] Martin, D. R.; Fowlkes, C. C.; Malik, J. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 26, No. 5, 530–549, 2004.

[62] Endres, I.; Hoiem, D. Category independent object proposals. In: *Computer Vision – ECCV 2010. Lecture Notes in Computer Science, Vol. 6315* Daniilidis, K.; Maragos, P.; Paragios, N. Eds. Springer Berlin Heidelberg, 575–588, 2010.

[63] Judd, T.; Ehinger, K.; Durand, F.; Torralba, A. Learning to predict where humans look. In: Proceedings of the IEEE 12th International Conference on Computer Vision, 2106–2113, 2009.

[64] Hou, X.; Zhang, L. Saliency detection: A spectral residual approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–8, 2007.

[65] Borji, A.; Itti, L. Exploiting local and global patch rarities for saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 478–485, 2012.

[66] Borji, A. Boosting bottom–up and top–down visual features for saliency estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 438–445, 2012.

[67] Viola, P.; Jones, M. Rapid object detection using a boosted cascade of simple features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2001.

[68] Felzenszwalb, P. F.; Girshick, B.; McAllester, D.; Ramanan, D. Object detection with discriminatively trained part-based models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 32, No. 9, 1627–1645, 2010.

[69] LeCun, Y.; Bottou, L.; Bengio, Y.; Haffner, P. Gradient-based learning applied to document recognition. *Proceedings of the IEEE* Vol. 86, No. 11, 2278–2324, 1998.

[70] Long, J.; Shelhamer, E.; Darrell, T. Fully convolutional networks for semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3431–3440, 2015.

[71] Hua, G.; Liu, Z. C.; Zhang, Z. Y.; Wu, Y. Iterative local-global energy minimization for automatic extraction of objects of interest. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 28, No. 10, 1701–1706, 2006.

[72] Ko, B. C.; Nam, J.-Y. Automatic object-of-interest segmentation from natural images. In: Proceedings of the 18th International Conference on Pattern Recognition, 45–48, 2006.

[73] Allili, M. S.; Ziou, D. Object of interest segmentation and tracking by using feature selection and active contours. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–8, 2007.

[74] Hu, Y.; Rajan, D.; Chia, L.-T. Robust subspace analysis for detecting visual attention regions in images. In: Proceedings of the 13th Annual ACM International Conference on Multimedia, 716–724, 2005.

[75] Vidal, R.; Ma, Y.; Sastry, S. Generalized principal component analysis (GPCA). *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 27, No. 12, 1945–1959, 2005.

[76] Rosin, P. L. A simple method for detecting salient regions. *Pattern Recognition* Vol. 42, No. 11, 2363–2371, 2009.

[77] Valenti, R.; Sebe, N.; Gevers, T. Image saliency by isocentric curvedness and color. In: Proceedings of the IEEE 12th International Conference on Computer Vision, 2185–2192, 2009.

[78] Klein, D. A.; Frintrop, S. Center-surround divergence of feature statistics for salient object detection. In: Proceedings of the International Conference on Computer Vision, 2214–2219, 2011.

[79] Li, X.; Li, Y.; Shen, C.; Dick, A. R.; van den Hengel, A. Contextual hypergraph modeling for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision, 3328–3335, 2013.

[80] Margolin, R.; Tal, A.; Zelnik-Manor, L. What makes a patch distinct? In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1139–1146, 2013.

[81] Felzenszwalb, P. F.; Huttenlocher, D. P. Efficient graph-based image segmentation. *International Journal of Computer Vision* Vol. 59, No. 2, 167–181, 2004.

[82] Levinshtein, A.; Stere, A.; Kutulakos, K. N.; Fleet, D. J.; Dickinson, S. J.; Siddiqi, K. TurboPixels: Fast superpixels using geometric flows. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 31, No. 12, 2290–2297, 2009.

[83] Yu, Z. W.; Wong, H. S. A rule based technique for extraction of visual attention regions based on real-time clustering. *IEEE Transactions on Multimedia* Vol. 9, No. 4, 766–784, 2007.

[84] Cheng, M. M.; Mitra, N. J.; Huang, X. L.; Torr, P. H. S.; Hu, S. M. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 37, No. 3, 569–582, 2015.

[85] Scharfenberger, C.; Wong, A.; Fergani, K.; Zelek, J. S.; Clausi, D. A. Statistical textural distinctiveness for salient region detection in natural images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 979–986, 2013.

[86] Cheng, M.-M.; Warrell, J.; Lin, W.-Y.; Zheng, S.; Vineet, V.; Crook, N. Efficient salient region detection with soft image abstraction. In: Proceedings of the IEEE International Conference on Computer Vision, 1529–1536, 2013.

[87] Jiang, Z.; Davis, L. S. Submodular salient region detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2043–2050, 2013.

[88] Adams, A.; Baek, J.; Davis, M. A. Fast high-dimensional filtering using the permutohedral lattice. *Computer Graphics Forum* Vol. 29, No. 2, 753–762, 2010.

[89] Shi, K. Y.; Wang, K. Z.; Lu, J. B.; Lin, L. PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2115–2122, 2013.

[90] Yu, H. N.; Li, J.; Tian, Y. H.; Huang, T. J. Automatic interesting object extraction from images using complementary saliency maps. In: Proceedings of the International Conference on Multimedia, 891–894, 2010.

[91] Lu, Y.; Zhang, W.; Lu, H.; Xue, X. Salient object detection using concavity context. In: Proceedings of the International Conference on Computer Vision, 233–240, 2011.

[92] Chang, K.-Y.; Liu, T.-L.; Chen, H.-T.; Lai, S.-H. Fusing generic objectness and visual saliency for salient object detection. In: Proceedings of the International Conference on Computer Vision, 914–921, 2011.

[93] Jiang, H.; Wang, J.; Yuan, Z.; Liu, T.; Zheng, N. Automatic salient object segmentation based on context and shape prior. In: Proceedings of the British Machine Vision Conference, 2011.

[94] Shen, X.; Wu, Y. A unified approach to salient object detection via low rank matrix recovery. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 853–860, 2012.

[95] Wei, Y. C.; Wen, F.; Zhu, W. J.; Sun, J. Geodesic saliency using background priors. In: *Computer Vision – ECCV 2012. Lecture Notes in Computer Science, Vol. 7574*. Fitzgibbon, A.; Lazebnik, S.; Perona, P.; Sato, Y.; Schmid, C. Eds. Springer Berlin Heidelberg, 29–42, 2012.

[96] Xie, Y. L.; Lu, H. C.; Yang, M. H. Bayesian saliency via low and mid level cues. *IEEE Transactions on Image Processing* Vol. 22, No. 5, 1689–1698, 2013.

[97] Yang, C.; Zhang, L.; Lu, H.; Ruan, X.; Yang, M.-H. Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3166–3173, 2013.

[98] Li, X.; Lu, H.; Zhang, L.; Ruan, X.; Yang, M.-H. Saliency detection via dense and sparse reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision, 2976–2983, 2013.

[99] Jiang, B.; Zhang, L.; Lu, H.; Yang, C.; Yang, M.-H. Saliency detection via absorbing Markov chain. In: Proceedings of the IEEE International Conference on Computer Vision, 1665–1672, 2013.

[100] Jiang, P.; Ling, H.; Yu, J.; Peng, J. Salient region detection by UFO: Uniqueness, focusness and objectness. In: Proceedings of the IEEE International Conference on Computer Vision, 1976–1983, 2013.

[101] Jia, Y.; Han, M. Category-independent object-level saliency detection. In: Proceedings of the IEEE International Conference on Computer Vision, 1761–1768, 2013.

[102] Zou, W.; Kpalma, K.; Liu, Z.; Ronsin, J. Segmentation driven low-rank matrix recovery for saliency detection. In: Proceedings of the 24th British Machine Vision Conference, 1–13, 2013.

[103] Peng, H.; Li, B.; Ji, R.; Hu, W.; Xiong, W.; Lang, C. Salient object detection via low-rank and structured sparse matrix decomposition. In: Proceedings of the 27th AAAI Conference on Artificial Intelligence, 796–802, 2013.

[104] Liu, R.; Cao, J.; Lin, Z.; Shan, S. Adaptive partial differential equation learning for visual saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3866–3873, 2014.

[105] Zhu, W.; Liang, S.; Wei, Y.; Sun, J. Saliency optimization from robust background detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2814–2821, 2014.

[106] Zhang, J.; Sclaroff, S. Saliency detection: A Boolean map approach. In: Proceedings of the IEEE International Conference on Computer Vision, 153–160, 2013.

[107] Li, N.; Ye, J.; Ji, Y.; Ling, H.; Yu, J. Saliency detection on light field. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2806–2813, 2014.

[108] Rahtu, E.; Kannala, J.; Salo, M.; Heikkilä, J. Segmenting salient objects from images and videos. In: *Computer Vision – ECCV 2010. Lecture Notes in Computer Science, Vol. 6315*. Daniilidis, K.; Maragos, P.; Paragios, N. Eds. Springer Berlin Heidelberg, 366–379, 2010.

[109] Khuwuthyakorn, P.; Robles-Kelly, A.; Zhou, J. Object of interest detection by saliency learning. In: *Computer Vision – ECCV 2010. Lecture Notes in Computer Science, Vol. 6312*. Daniilidis, K.; Maragos, P.; Paragios, N. Eds. Springer Berlin Heidelberg, 636–649, 2010.

[110] Mehrani, P.; Veksler, O. Saliency segmentation based on learning and graph cut refinement. In: Proceedings of the British Machine Vision Conference, 110.1–110.12. 2010.

[111] Lu, S.; Mahadevan, V.; Vasconcelos, N. Learning optimal seeds for diffusion-based salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2790–2797, 2014.

[112] Kim, J.; Han, D.; Tai, Y.-W.; Kim, J. Salient region detection via high-dimensional color transform. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 883–890, 2014.

[113] Marchesotti, L.; Cifarelli, C.; Csurka, G. A framework for visual saliency detection with applications to image thumbnailing. In: Proceedings of the IEEE 12th International Conference on Computer Vision, 2232–2239, 2009.

[114] Wang, M.; Konrad, J.; Ishwar, P.; Jing, K.; Rowley, H. Image saliency: From intrinsic to extrinsic context. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 417–424, 2011.

[115] Mai, L.; Niu, Y.; Liu, F. Saliency aggregation: A datadriven approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1131–1138, 2013.

[116] Zhai, Y.; Shah, M. Visual attention detection in video sequences using spatiotemporal cues. In: Proceedings of the 14th ACM International Conference on Multimedia, 815–824, 2006.

[117] Liu, T.; Zheng, N.; Ding, W.; Yuan, Z. Video attention: Learning to detect a salient object sequence. In: Proceedings of the 19th International Conference on Pattern Recognition, 1–4, 2008.

[118] Li, Y.; Sheng, B.; Ma, L. Z.; Wu, W.; Xie, Z. F. Temporally coherent video saliency using regional dynamic contrast. *IEEE Transactions on Circuits and Systems for Video Technology* Vol. 23, No. 12, 2067–2076, 2013.

[119] Li, H. L.; Ngan, K. N. A co-saliency model of image pairs. *IEEE Transactions on Image Processing* Vol. 20, No. 12, 3365–3375, 2011.

[120] Chang, K.; Liu, T.; Lai, S. From co-saliency to co-segmentation: An efficient and fully unsupervised energy minimization model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2129–2136, 2011.

[121] Fu, H. Z.; Cao, X. C.; Tu, Z. W. Cluster-based co-saliency detection. *IEEE Transactions on Image Processing* Vol. 22, No. 10, 3766–3778, 2013.

[122] Niu, Y.; Geng, Y.; Li, X.; Liu, F. Leveraging stereopsis for saliency analysis. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 454–461, 2012.

[123] Desingh, K.; Krishna, K. M.; Rajan, D.; Jawahar, C. V. Depth really matters: Improving visual salient region detection with depth. In: Proceedings of the British Machine Vision Conference, 2013.

[124] Rother, C.; Minka, T. P.; Blake, A.; Kolmogorov, V. Cosegmentation of image pairs by histogram matching-incorporating a global constraint into MRFs. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 993–1000, 2006.

[125] Batra, D.; Kowdle, A.; Parikh, D.; Luo J.; Chen, T. iCoseg: Interactive co-segmentation with intelligent scribble guidance. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 3169–3176, 2010.

[126] Mukherjee, L.; Singh, V.; Peng, J. Scale invariant cosegmentation for image groups. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1881–1888, 2011.

[127] Kim, G.; Xing, E. P.; Li, F. F.; Kanade, T. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In: Proceedings of the International Conference on Computer Vision Barcelona, 169–176, 2011.

[128] Feng, J.; Wei, Y.; Tao, L.; Zhang, C.; Sun, J. Salient object detection by composition. In: Proceedings of the International Conference on Computer Vision, 1028–1035, 2011.

[129] Wang, P.; Wang, J.; Zeng, G.; Feng, J.; Zha, H.; Li, S. Salient object detection for searched web images via global saliency. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Providence, 3194–3201, 2012.

[130] Wang, L.; Xue, J.; Zheng, N.; Hua, G. Automatic salient object extraction with contextual cue. In: Proceedings of the International Conference on Computer Vision, 105–112, 2011.

[131] Tian, Y. H.; Li, J.; Yu, S.; Huang, T. J. Learning complementary saliency priors for foreground object segmentation in complex scenes. *International Journal of Computer Vision* Vol. 111, No. 2, 153–170, 2015.

[132] Borji, A.; Cheng, M. M.; Jiang, H. Z.; Li, J. Salient object detection: A benchmark. *IEEE Transactions on Image Processing* Vol. 24, No. 12, 5706–5722, 2015.

[133] Li, J.; Tian, Y. H.; Duan, L. Y.; Huang, T. J. Estimating visual saliency through single image optimization. *IEEE Signal Processing Letters* Vol. 20, No. 9, 845–848, 2013.

[134] Shi, J. B.; Malik, J. Normalized cuts and image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 22, No. 8, 888–905, 2000.

[135] Tu, Z. W.; Bai, X. Auto-context and its application to high-level vision tasks and 3D brain image segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 32, No. 10, 1744–1757, 2010.

[136] Qin, Y.; Lu, H. C.; Xu, Y. Q.; Wang, H. Saliency detection via cellular automata. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 110–119, 2015.

[137] Li, Y.; Sun, J.; Tang, C. K.; Shum, H. Y. Lazy snapping. *ACM Transactions on Graphics* Vol. 23, No. 3, 303–308, 2004.

[138] Rother, C.; Kolmogorov, V.; Blake, A. "GrabCut": Interactive foreground extraction using iterated graph cuts. *ACM Transactions on Graphics* Vol. 23, No. 3, 309–314, 2004.

[139] Lang, C. Y.; Nguyen, T. V.; Katti, H.; Yadati, K.; Kankanhalli, M.; Yan, S. C. Depth matters: Influence of depth cues on visual saliency. In: *Computer Vision – ECCV 2012. Lecture Notes in Computer Science, Vol. 7573*. Fitzgibbon, A.; Lazebnik, S.; Perona, P.; Sato, Y.; Schmid, C. Eds. Springer Berlin Heidelberg, 101–115, 2012.

[140] Zhang, J.; Wang, M.; Lin, L.; Yang, X.; Gao, J.; Rui, Y. Saliency detection on light field. *ACM Transactions on Multimedia Computing, Communications, and Applications* Vol. 13, No. 3, 1–22, 2017.

[141] Krizhevsky, A.; Sutskever, I.; Hinton, G. E. ImageNet classification with deep convolutional neural networks. *Communications of the ACM* Vol. 60, No. 6, 84–90, 2017.

[142] Xie, S.; Tu, Z. Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision, 1395–1403, 2015.

[143] Szegedy, C.; Liu, W.; Jia, Y.; Sermanet, P.; Reed, S.; Anguelov, D.; Erhan, D.; Vanhoucke, V.; Rabinovich, A. Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–9, 2015.

[144] Lee, G.; Tai, Y.-W.; Kim, J. Deep saliency with encoded low level distance map and high level features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 660–668, 2016.

[145] Kim, J.; Pavlovic, V. A shape preserving approach for salient object detection using convolutional neural networks. In: Proceedings of the 23rd International Conference on Pattern Recognition, 609–614, 2016.

[146] Wang, X.; Ma, H. M.; Chen, X. Z. Salient object detection via fast R-CNN and low-level cues. In: Proceedings of the IEEE International Conference on Image Processing, 1042–1046, 2016.

[147] Kim, J.; Pavlovic, V. A shape preserving approach for salient object detection using convolutional neural networks. In: Proceedings of the 23rd International Conference on Pattern Recognition, 609–614, 2016.

[148] Chen, T. S.; Lin, L.; Liu, L. B.; Luo, X. N.; Li, X. L. DISC: Deep image saliency computing via progressive representation learning. *IEEE Transactions on Neural Networks and Learning Systems* Vol. 27, No. 6, 1135–1149, 2016.

[149] Li, H. Y.; Chen, J.; Lu, H. C.; Chi, Z. Z. CNN for saliency detection with low-level feature integration. *Neurocomputing* Vol. 226, 212–220, 2017.

[150] Liu, N.; Han, J. DHSNet: Deep hierarchical saliency network for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 678–686, 2016.

[151] Li, G. B.; Yu, Y. Z. Deep contrast learning for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 478–487, 2016.

[152] Simonyan, K.; Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint* arXiv:1409.1556, 2014.

[153] Kuen, J.; Wang, Z. H.; Wang, G. Recurrent attentional networks for saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3668–3677, 2016.

[154] Kruthiventi, S. S. S.; Gudisa, V.; Dholakiya, J. H.; Babu, R. V. Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5781–5790, 2016.

[155] Tang, Y. B.; Wu, X. Q. Saliency detection via combining region-level and pixel-level predictions with CNNs. In: *Computer Vision – ECCV 2016. Lecture Notes in Computer Science, Vol. 9912.* Leibe, B.; Matas, J.; Sebe, N.; Welling, M. Eds. Springer Cham, 809–825 2016.

[156] Tang, Y. B.; Wu, X. Q.; Bu, W. Deeply-supervised recurrent convolutional neural network for saliency detection. In: Proceedings of the ACM on Multimedia Conference, 397–401, 2016.

[157] Li, X.; Zhao, L. M.; Wei, L. N.; Yang, M. H.; Wu, F.; Zhuang, Y. T. et al. DeepSaliency: Multi-task deep neural network model for salient object detection. *IEEE Transactions on Image Processing* Vol. 25, No. 8, 3919–3930, 2016.

[158] Zhang, J.; Dai, Y. C.; Porikli, F. Deep salient object detection by integrating multi-level cues. In: Proceedings of the IEEE Winter Conference on Applications of Computer Vision, 1–10, 2017.

[159] Li, G.; Xie, Y.; Lin, L.; Yu, Y. Instance-level salient object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2386–2395, 2017.

[160] Wang, L. J.; Lu, H. C.; Ruan, X.; Yang, M. H. Deep networks for saliency detection via local estimation and global search. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3183–3192, 2015.

[161] Girshick, R.; Donahue, J.; Darrell, T.; Malik, J. Rich feature hierarchies for accurate object detection and semantic segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 580–587, 2014.

[162] Uijlings, J. R. R.; van de Sande, K. E. A.; Gevers, T.; Smeulders, A. W. M. Selective search for object recognition. *International Journal of Computer Vision* Vol. 104, No. 2, 154–171, 2013.

[163] Zeiler, M. D.; Fergus, R. Visualizing and understanding convolutional networks. In: *Computer Vision – ECCV 2014. Lecture Notes in Computer Science, Vol. 8689.* Fleet, D.; Pajdla, T.; Schiele, B.; Tuytelaars, T. Eds. Springer Cham, 818–833, 2014.

[164] Lee, C.-Y.; Xie, S.; Gallagher, P.; Zhang, Z.; Tu, Z. Deeply-supervised nets. In: Proceedings of the 18th International Conference on Artificial Intelligence and Statistics, 562–570, 2015.

[165] Jaderberg, M.; Simonyan, K.; Zisserman, A.; Kavukcuoglu, K. Spatial transformer networks. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, Vol. 2, 2017–2025, 2015.

[166] Achanta, R.; Shaji, A.; Smith, K.; Lucchi, A.; Fua, P.; Süsstrunk, S. SLIC superpixels compared to state-of-the-art superpixel methods. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 34, No. 11, 2274–2282, 2012.

[167] Chen, L. C.; Papandreou, G.; Kokkinos, I.; Murphy, K.; Yuille, A. L. DeepLab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 40, No. 4, 834–848, 2018.

[168] Arbelaez, P.; PontTuset, J.; Barron, J. T.; Marques, F.; Malik, J. Multiscale combinatorial grouping. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 328–335, 2014.

[169] Krähenbühl, P.; Koltun, V. Efficient inference in fully connected CRFs with Gaussian edge potentials. In: Proceedings of the 24th International Conference on Neural Information Processing Systems, 109–117, 2011.

[170] Russakovsky, O.; Deng, J.; Su, H.; Krause, J.; Satheesh, S.; Ma, S. A.; Huang, Z.; Karpathy, A.; Khosla, A.; Bernstein, M. et al. ImageNet large scale visual recognition challenge. *International Journal of Computer Vision* Vol. 115, No. 3, 211–252, 2015.

[171] Oquab, M.; Bottou, L.; Laptev, I.; Sivic, J. Learning and transferring mid-level image representations using convolutional neural networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1717–1724, 2014.

[172] Jiang, H.; Wang, J.; Yuan, Z.; Wu, Y.; Zheng, N.; Li, S. Salient object detection: A discriminative regional feature integration approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2083–2090, 2013.

[173] Zhang, G. X.; Cheng, M. M.; Hu, S. M.; Martin, R. R. A shape-preserving approach to image resizing. *Computer Graphics Forum* Vol. 28, No. 7, 1897–1906, 2009.

[174] Huang, H.; Zhang, L.; Zhang, H. C. Arcimboldo-like collage using Internet images. *ACM Transactions on Graphics* Vol. 30, No. 6, Article No. 155, 2011.

[175] Liu, H.; Zhang, L.; Huang, H. Web-image driven best views of 3D shapes. *The Visual Computer* Vol. 28, No. 3, 279–287, 2012.

[176] Zhu, J.; Wu, J.; Wei, Y.; Chang, E.; Tu, Z. Unsupervised object class discovery via saliency-guided multiple class learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3218–3225, 2012.

[177] Chen, T.; Cheng, M.-M.; Tan, P.; Shamir, A.; Hu, S.-M. Sketch2Photo: Internet image montage. *ACM Transactions on Graphics* Vol. 28, No. 5, Article No. 124, 2009.

[178] Goldberg, C.; Chen, T.; Zhang, F. L.; Shamir, A.; Hu, S. M. Data-driven object manipulation in images. *Computer Graphics Forum* Vol. 31, No. 2pt1, 265–274, 2012.

[179] Chia, A. Y.-S.; Zhuo, S.; Gupta, R. K.; Tai, Y.-W.; Cho, S.-Y.; Tan, P.; Lin, S. Semantic colorization with internet images. *ACM Transactions on Graphics* Vol. 30, No. 6, Article No. 156, 2011.

[180] Rutishauser, U.; Walther, D.; Koch, C.; Perona P. Is bottom–up attention useful for object recognition? In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, II, 2004.

[181] Kanan, C.; Cottrell, G. Robust classification of objects, faces, and flowers using natural image statistics. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2472–2479, 2010.

[182] Moosmann, F.; Larlus, D.; Jurie, F. Learning saliency maps for object categorization. In: Proceedings of the International Workshop on the Representation and Use of Prior Knowledge in Vision, 2006.

[183] Borji, A.; Ahmadabadi, M. N.; Araabi, B. N. Cost-sensitive learning of top-down modulation for attentional control. *Machine Vision and Applications* Vol. 22, No. 1, 61–76, 2011.

[184] Borji, A.; Itti, L. Scene classification with a sparse set of salient regions. In: Proceedings of the IEEE International Conference on Robotics and Automation, 1902–1908, 2011.

[185] Shen, H.; Li, S. X.; Zhu, C. F.; Chang, H. X.; Zhang, J. L. Moving object detection in aerial video based on spatiotemporal saliency. *Chinese Journal of Aeronautics* Vol. 26, No. 5, 1211–1217, 2013.

[186] Ren, Z. X.; Gao, S. H.; Chia, L. T.; Tsang, I. W. H. Region-based saliency detection and its application in object recognition. *IEEE Transactions on Circuits and Systems for Video Technology* Vol. 24, No. 5, 769–779, 2014.

[187] Guo, C. L.; Zhang, L. M. A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression. *IEEE*

Transactions on Image Processing Vol. 19, No. 1, 185–198, 2010.

[188] Itti, L. Automatic foveation for video compression using a neurobiological model of visual attention. *IEEE Transactions on Image Processing* Vol. 13, No. 10, 1304–1318, 2004.

[189] Ma, Y. F.; Hua, X. S.; Lu, L.; Zhang, H. J. A generic framework of user attention model and its application in video summarization. *IEEE Transactions on Multimedia* Vol. 7, No. 5, 907–919, 2005.

[190] Lee, Y. J.; Ghosh, J.; Grauman, K. Discovering important people and objects for egocentric video summarization. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1346–1353, 2012.

[191] Ji, Q. G.; Fang, Z. D.; Xie, Z. H.; Lu, Z. M. Video abstraction based on the visual attention model and online clustering. *Signal Processing: Image Communication* Vol. 28, No. 3, 241–253, 2013.

[192] Goferman, S.; Tal, A.; Zelnik-Manor, L. Puzzle-like collage. *Computer Graphics Forum* Vol. 29, No. 2, 459–468, 2010.

[193] Wang, J.; Quan, L.; Sun, J.; Tang, X.; Shum, H.-Y. Picture collage. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 347–354, 2006.

[194] Ninassi, A.; Le Meur, O.; Le Callet, P.; Barba, D. Does where you gaze on an image affect your perception of quality? Applying visual attention to image quality metric. In: Proceedings of the IEEE International Conference on Image Processing, II169–II172, 2007.

[195] Liu, H. T.; Heynderickx, I. Studying the added value of visual attention in objective image quality metrics based on eye movement data. In: Proceedings of the 16th IEEE International Conference on Image Processing, 3097–3100, 2009.

[196] Li, A.; She, X.; Sun, Q. Color image quality assessment combining saliency and FSIM. In: Proceedings of the SPIE 8878, 5th International Conference on Digital Image Processing, 88780I, 2013.

[197] Donoser, M.; Urschler, M.; Hirzer, M.; Bischof, H. Saliency driven total variation segmentation. In: Proceedings of the IEEE 12th International Conference on Computer Vision, 817–824, 2009.

[198] Li, Q.; Zhou, Y.; Yang, J. Saliency based image segmentation. In: Proceedings of the International Conference on Multimedia Technology, 5068–5071, 2011.

[199] Qin, C. C.; Zhang, G. P.; Zhou, Y. C.; Tao, W. B.; Cao, Z. G. Integration of the saliency-based seed extraction and random walks for image segmentation. *Neurocomputing* Vol. 129, 378–391, 2014.

[200] Johnson-Roberson, M.; Bohg, J.; Björkman, M.; Kragic, D. Attention-based active 3D point cloud segmentation. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 1165–1170, 2010.

[201] Feng, S. H.; Xu, D.; Yang, X. Attention-driven salient edge(s) and region(s) extraction with application to CBIR. *Signal Processing* Vol. 90, No. 1, 1–15, 2010.

[202] Sun, J. D.; Xie, J. C.; Liu, J.; Sikora, T. Image adaptation and dynamic browsing based on two-layer saliency combination. *IEEE Transactions on Broadcasting* Vol. 59, No. 4, 602–613, 2013.

[203] Li, L.; Jiang, S. Q.; Zha, Z. J.; Wu, Z. P.; Huang, Q. M. Partial-duplicate image retrieval via saliency-guided visual matching. *IEEE MultiMedia* Vol. 20, No. 3, 13–23, 2013.

[204] Stalder, S.; Grabner, H.; Van Gool, L. Dynamic objectness for adaptive tracking. In: *Computer Vision – ACCV 2012. Lecture Notes in Computer Science, Vol. 7726.* Lee, K. M.; Matsushita, Y.; Rehg, J. M.; Hu, Z. Eds. Springer Berlin Heidelberg, 43–56, 2013.

[205] Li, J.; Levine, M. D.; An, X. J.; Xu, X.; He, H. G. Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 35, No. 4, 996–1010, 2013.

[206] García, G. M.; Klein, D. A.; Stückler, J.; Frintrop, S.; Cremers, A. B. Adaptive multi-cue 3D tracking of arbitrary objects. In: *Pattern Recognition. Lecture Notes in Computer Science, Vol. 7476.* Pinz, A.; Pock, T.; Bischof, H.; Leberl, F. Eds. Springer Berlin Heidelberg, 357–366, 2012.

[207] Borji, A.; Frintrop, S.; Sihite, D. N.; Itti, L. Adaptive object tracking by learning background context. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 23–30, 2012.

[208] Klein, D. A.; Schulz, D.; Frintrop, S.; Cremers, A. B. Adaptive real-time video-tracking for arbitrary objects. In: Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, 772–777, 2010.

[209] Frintrop, S.; Kessel, M. Most salient region tracking. In: Proceedings of the IEEE International Conference on Robotics and Automation, 1869–1874, 2009.

[210] Zhang, G.; Yuan, Z.; Zheng, N.; Sheng, X.; Liu, T. Visual saliency based object tracking. In: *Computer Vision – ACCV 2009. Lecture Notes in Computer Science, Vol. 5995.* Zha, H.; Taniguchi, R.; Maybank, S. Eds. Springer Berlin Heidelberg, 193–203, 2010.

[211] Karpathy, A.; Miller, S.; Li, F-F. Object discovery in 3D scenes via shape analysis. In: Proceedings of the IEEE International Conference on Robotics and Automation, 2088–2095, 2013.

[212] Frintrop, S.; Garcıa, G. M.; Cremers, A. B. A cognitive approach for object discovery. In: Proceedings of the 22nd International Conference on Pattern Recognition, Stockholm, 2329–2334, 2014.

[213] Meger, D.; Forssén, P. E.; Lai, K.; Helmer, S.; McCann, S.; Southey, T.; Baumann, M.; Little, J. J.; Lowe, D. G. Curious George: An attentive semantic robot. *Robotics and Autonomous Systems* Vol. 56, No. 6, 503–511, 2008.

[214] Sugano, Y.; Matsushita, Y.; Sato, Y. Calibration-free gaze sensing using saliency maps. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2667–2674, 2010.

[215] Msra dataset. Available at `http://research.microsoft.com/en-us/um/people/jiansun/`.

[216] Alpert, S.; Galun, M.; Basri, R.; Brandt, A. Image segmentation by probabilistic bottom–up aggregation and cue integration. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1–8, 2007.

[217] Movahedi, V.; Elder, J. H. Design and perceptual validation of performance measures for salient object segmentation. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, 49–56, 2010.

[218] Brown, M.; Süsstrunk, S. Multi-spectral SIFT for scene category recognition. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition, 177–184, 2011.

[219] Wang, Q.; Yan, P. K.; Yuan, Y.; Li, X. L. Multi-spectral saliency detection. *Pattern Recognition Letters* Vol. 34, No. 1, 34–41, 2013.

[220] Msra10k dataset. Available at `http://www.cse.cuhk.edu.hk/leojia/projects/hsaliency/`.

[221] Thur15k dataset. Available at `http://mmcheng.net/gsal/`.

[222] Judd dataset. Available at `http://ilab.usc.edu/borji/Resources.html`.

[223] Koehler, K.; Guo, F.; Zhang, S.; Eckstein, M. P. What do saliency models predict? *Journal of Vision* Vol. 14, No. 3, 14, 2014.

[224] Xu, J.; Jiang, M.; Wang, S.; Kankanhalli, M. S.; Zhao, Q. Predicting human gaze beyond pixels. *Journal of Vision* Vol. 14, No. 1, 28, 2014.

[225] Li, J.; Tian, Y.; Huang, T.; Gao, W. A dataset and evaluation methodology for visual saliency in video. In: Proceedings of the IEEE International Conference on Multimedia and Expo, 442–445, 2009.

[226] Wu, Y.; Zheng, N. N.; Yuan, Z. J.; Jiang, H. Z.; Liu, T. Detection of salient objects with focused attention based on spatial and temporal coherence. *Chinese Science Bulletin* Vol. 56, No. 10, 1055–1062, 2011.

[227] Avidan, S.; Shamir, A. Seam carving for content-aware image resizing. *ACM Transactions on Graphics* Vol. 26, No. 3, Article No. 10, 2007.

[228] Borji, A.; Cheng, M. M.; Jiang, H. Z.; Li, J. Salient object detection: A benchmark. *IEEE Transactions on Image Processing* Vol. 24, No. 12, 5706–5722, 2015.

[229] He, K.; Zhang, X.; Ren, S.; Sun J. Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 770–778, 2016.

[230] Torralba A.; Efros, A. A. Unbiased look at dataset bias. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1521–1528, 2011.

[231] Tatler, B. W. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision* Vol. 7, No. 14, 4, 2007.

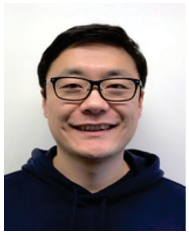
[232] Borji, A.; Sihite, D. N.; Itti, L. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing* Vol. 22, No. 1, 55–69, 2013.

[233] Cheng, M. M.; Mitra, N. J.; Huang, X. L.; Hu, S. M. SalientShape: Group saliency in image collections. *The Visual Computer* Vol. 30, No. 4, 443–453, 2014.

Ali Borji received his Ph.D. degree in cognitive neuronsciences from the Institute for Studies in Fundamental Sciences (IPM), Tehran, Iran, 2009. He is currently a senior research scientist at MarkableAI, New York, USA. His research interests include visual attention, visual search, and object and scene recognition.

Ming-Ming Cheng received his Ph.D. degree from Tsinghua University in 2012. Then he did two-year research fellow, with Prof. Philip Torr in Oxford. He is now a full professor at Nankai University, leading the Media Computing Lab. His research interests include computer graphics, computer vision, and image processing. He received research awards including ACM China Rising Star Award, IBM Global SUR Award, CCF-Intel Young Faculty Researcher Program, etc.

Qibin Hou is currently a Ph.D. candidate with College of Computer Science and Control Engineering, Nankai University, under the supervision of Prof. Ming-Ming Cheng. His research interests include deep learning, image processing, and computer vision.



Huaizu Jiang is a Ph.D. candidate at College of Information and Computer Sciences, University of Massachusetts, Amherst, USA. His research interests include computer vision, computational photography, machine learning, natural language processing, and artificial intelligence in general. His Ph.D. research is about large-scale visual learning from unlabeled data, particularly from unlabeled videos. He obtained his M.E. and B.E. degrees from Xi'an Jiaotong University, China, in 2009 and 2012, respectively.

Jia Li is currently an associate professor with the School of Computer Science and Engineering, Beihang University, Beijing, China. He received his B.E. degree from Tsinghua University in Jul. 2005 and Ph.D. degree from the Institute of Computing Technology, Chinese Academy of Sciences, in Jan. 2011. Before he joined Beihang University, he used to serve in Nanyang Technological University, Peking University, and Shanda Innovations. His research interests include computer vision and multimedia big data, especially the cognitive vision towards evolvable algorithms and models. He is the author or coauthor of over 60 technical articles in refereed journals and conferences such as TPAMI, IJCV, TIP, ICCV, and CVPR. More information can be found at http://cvteam.net.