Salient Objects in Clutter

Deng-Ping Fan, Jing Zhang, Gang Xu, Ming-Ming Cheng, and Ling Shao, *Fellow, IEEE*

Abstract—In this paper, we identify and address a serious design bias of existing salient object detection (SOD) datasets, which unrealistically assume that each image should contain at least one clear and uncluttered salient object. This *design bias* has led to a saturation in performance for state-of-the-art SOD models when evaluated on existing datasets. However, these models are still far from satisfactory when applied to real-world scenes. Based on our analyses, we propose a new high-quality dataset and update the previous saliency benchmark. Specifically, our dataset, called Salient Objects in Clutter **(SOC)**, includes images with both salient and non-salient objects from several common object categories. In addition to object category annotations, each salient image is accompanied by attributes that reflect common challenges in common scenes, which can help provide deeper insight into the SOD problem. Further, with a given saliency encoder, *e.g.*, the backbone network, existing saliency models are designed to achieve mapping from the training image set to the training ground-truth set. We therefore argue that improving the dataset can yield higher performance gains than focusing only on the decoder design. With this in mind, we investigate several dataset-enhancement strategies, including label smoothing to implicitly emphasize salient boundaries, random image augmentation to adapt saliency models to various scenarios, and self-supervised learning as a regularization strategy to learn from small datasets. Our extensive results demonstrate the effectiveness of these tricks. We also provide a comprehensive benchmark for SOD, which can be found in our repository: https://github.com/DengPingFan/SODBenchmark.

Index Terms—Salient object detection, SOD, SOC, survey, dataset, benchmark.

1 Introduction

THIS paper considers the task of salient object detection (SOD), which aims to detect the most attention-grabbing objects in a scene and then extract pixel-accurate silhouettes for them. The merit of SOD lies in its many applications, including foreground map evaluation [1], [2], [3], visual tracking [4], [5], [6], action recognition [7], image retrieval [8], [9], information discovery [10], [11], image contrast enhancement [12], person re-identification [13] image segmentation [14], [15], video segmentation [16], photo synthesis [17], content-aware image editing [18], image caption [19], and video compression [20], [21], style transfer [22], [23], image matching [24], autonomous underwater robots [25], camouflaged object detection [26], [27], aesthetic scoring [28], self-driving vehicles [29], plant species identification [30], dichotomous image segmentation[1], VR/AR [31][2], Sony's BRAVIA XR TV[3], *etc*. However, existing SOD datasets [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42] are flawed either in their data collection procedure or data quality. Specifically, most datasets assume that an image should contain at least *one salient object, and thus they discard images that do not contain any salient objects*. We call this ***data selection bias [43].***

Moreover, existing datasets typically contain images with a single object or several uncluttered objects. These datasets do not adequately reflect the complexity of real-world images, where scenes usually contain multiple objects amidst significant clutter. As a result, all top-performing models trained on the existing large-scale datasets (*e.g.*, DUTS [41]) have nearly saturated performance (*e.g.*, SCRN [44] has an *S-measure* > 0.9 on ECC [37]), but still achieve unsatisfactory results on realistic images (*e.g.*, *S-measure* < 0.8 on *SOC* [45]). As the current SOD models are biased towards ideal conditions, their effectiveness may be impaired once they are applied to real-world scenes. To solve this problem, it is important to introduce a dataset with more realistic conditions.

Another issue faced by the RGB SOD community is that only the overall performance of the models can be analyzed using existing datasets. This is because none of the datasets contain attributes that reflect different challenges. Having such attributes

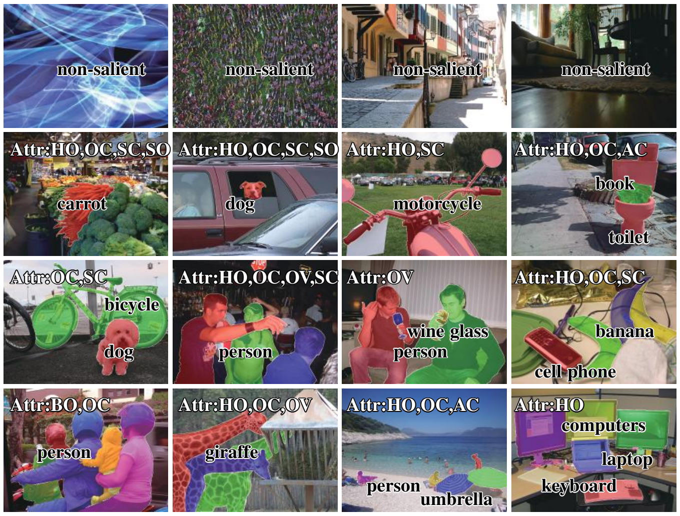


Figure 1. Examples from our new SOC dataset, including *non-salient* (first row) and *salient* object images (rows 2 to 4). For salient object images, an instance-level ground-truth map (different color), object attributes (Attr) and category labels are provided.

- *Deng-Ping Fan, Gang Xu and Ming-Ming Cheng are with the CS, Nankai University, Tianjin, China. (E-mail: dengpfan@gmail.com; gangxu@mail.nankai.edu.cn; cmm@nankai.edu.cn)*
- *Jing Zhang is with Research School of Engineering, the Australian National University, ACRV, DATA61-CSIRO. (Email: zjnwpu@gmail.com)*
- *Ling Shao is with the Inception Institute of Artificial Intelligence, Abu Dhabi, UAE. (E-mail: ling.shao@ieee.org)*
- *The major part of this work was done in Nankai University.*
- *Ming-Ming Cheng is the corresponding author.*
- *This work is funded by the National Key Research and Development Program of China (No.2018AAA0100400) and NSFC (NO. 61922046).*

[1] https://xuebinqin.github.io/dis/index.html
[2] AR CUT & PASTE: https://www.youtube.com/watch?v=VxJmS8avjbY.
[3] https://www.youtube.com/watch?v=4LnCuTAlVno&feature=youtu.be.

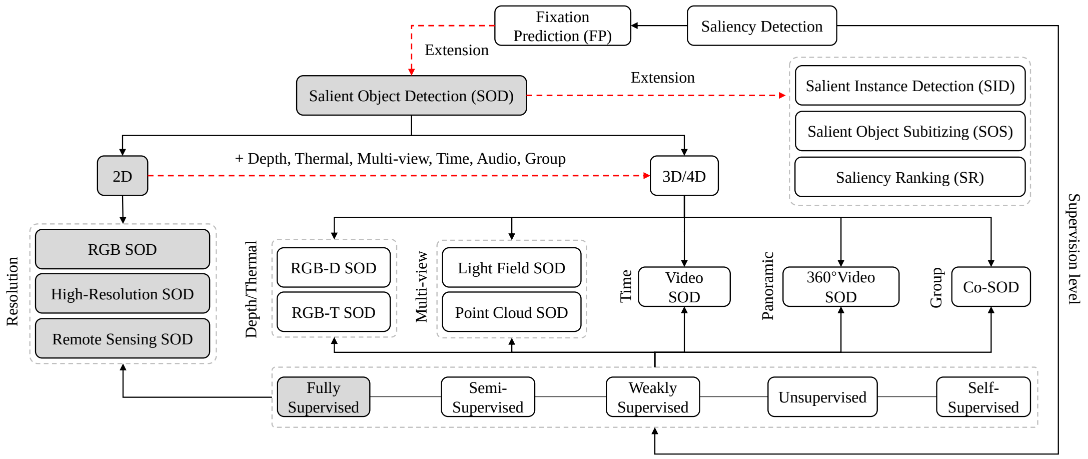


Figure 2. Taxonomy of the saliency detection task. We highlight the scope of this study in gray. See § 2 for details.

would help i) provide deeper insight into the SOD problem, ii) enable the pros and cons of the SOD models to be investigated, and iii) allow the model performances to be objectively assessed from different perspectives. Finally, with a given saliency encoder, *e.g.*, the backbone network, existing saliency models are designed to achieve mapping from the training image set to the training ground-truth set. We thus argue that efforts on improving the dataset, *e.g.*, fixing the data bias issue, can yield higher performance gains than focusing only on the decoder design. Towards this, we investigate several dataset-enhancement strategies, including label smoothing to highlight salient boundaries, random image augmentation to adapt saliency models to various scenarios, and self-supervised learning as a form of regularization to learn from small datasets. Extensive experiments validate the effectiveness of these tricks.

Our contributions are summarized as follows:

1) **Dataset.** We collect a new high-quality SOD dataset, named “Salient Objects in Clutter,” or **SOC**. *SOC* is the largest instance-level SOD dataset to date, containing 6,000 images from more than 80 common categories. It differs from existing datasets in three aspects: i) Salient objects have category annotations, which can be used for new research problems, such as weakly supervised SOD. ii) The inclusion of non-salient images and objects makes this dataset more realistic and challenging than the existing ones. iii) Salient objects have attributes that reflect various situations encountered in the real world, such as *motion blur*, *occlusion* and *background clutter*. As a consequence, *SOC* narrows the gap between existing datasets and real-world scenes.
2) **Review & Benchmark.** We present the largest scale RGB SOD study, reviewing **203** representative models including **84** algorithms using handcrafted features and **119** deep learning based models. Besides, we also maintain an online benchmark (*i.e.*, https://github.com/DengPingFan/SODBenchmark.) to dynamically trace the development of this field. In addition, we provide the most comprehensive benchmark of the **100 representative** SOD models. To evaluate the models, for the first time, we not only present the overall but also an attribute performance evaluation. This allows a deeper understanding of the models and provides a more complete benchmark.
3) **Strategy.** We investigate the biased dataset issue and introduce three dataset-enhancement strategies; namely, label smoothing to make the model aware of the salient boundaries, random image augmentation to adapt the saliency models to various common scenarios, and self-supervised learning as a regularization technique to learn from small datasets. Despite the apparent simplicity of our strategies, we can achieve an average absolute improvement of 1.14% S_α over nine existing cutting-edge models.
4) **Discussions & Future Directions.** Based on our *SOC*, we present the pros and cons of the current SOD algorithms, discuss several under-investigated open issues, and provide potential future directions at six levels, *e.g.*, the dataset level, task level, model level, supervision level, evaluation level, and application level.

This work extends our previous conference version [45] in the following aspects. First, we provide more details on our *SOC*, including sample images without salient objects, images with attributes, and statistics of the attributes. Second, we study three novel training dataset related strategies to fully utilize the non-salient object data and achieve the new state-of-the-art performance. Third, we conduct the largest-scale (46 traditional and 54 deep learning models) benchmarking of SOD models on our *SOC*. Finally, based on our benchmarking results, we highlight several fundamental research directions and challenges in the SOD.

2 Related Work

2.1 Scope

Salient object detection originated from the task of fixation prediction (FP) [53], [54], switching attention regions for accurate object-level regions. SOD can be traced back to the seminal works [55], [56]. Current algorithms have been developed for 2D images of limited resolution (width or height $<$ 500 pixels), high-resolution (*i.e.*, 1080p, 4K) [57], [58] and even remote sensing data [59]. According to the supervision strategy, there are five types of SOD models: fully supervised [60], semi-supervised [61], weakly supervised [62], [63], [64], unsupervised [65], [66], [67], and self-supervised [67], [68].

Recently, several interesting extensions of SOD have also been introduced, such as salient instance detection (SID) [51], [69], salient object subitizing (SOS) [49], [70], [71], and saliency ranking [72], [73]. A taxonomy of the saliency detection task is shown in Fig. 2. Different from previous SOD reviews [46], [74], [75], [76], [77], [78], [79], [80], [81], we mainly focus on 2D salient object detection in a fully supervised manner. We highlight the scope of this study in gray. For other closely related 3D/4D SOD tasks, we refer readers to recent survey and benchmarking works

Table 1
Summary of popular SOD datasets. Our SOC is the only one meeting all requirements. According to [46], these datasets can be grouped into three types: early (▲), popular/modern (♦), and special (◊). See § 2.2 for more details.

#	Dataset	Year	Publ.	High-Quality	$\geq 5k$	Non-Salient	Attribute	Category	Bounding Box	Object	Instance
1	MSRA-A, -B [32] ▲	2007	CVPR	✓	✓	-	-	-	✓	✓	-
2	SED1, SED2 [33] ▲	2007	CVPR	✓	-	-	-	-	-	✓	-
3	ASD [47] ▲	2009	CVPR	✓	-	-	-	-	-	✓	-
4	SOD [48] ♦	2010	CVPRW	✓	-	-	-	-	-	✓	-
5	M10K [35] ♦	2011	CVPR	✓	✓	-	-	-	-	✓	-
6	Judd-A [36] ▲	2012	ECCV	✓	-	-	-	-	-	✓	-
7	DU-O [38] ♦	2013	CVPR	✓	✓	-	-	-	✓	✓	-
8	ECC [37] ♦	2013	CVPR	✓	-	-	-	-	-	✓	-
9	PASCAL-S [39] ♦	2014	CVPR	✓	-	-	-	-	-	✓	-
10	HKU [40] ♦	2015	CVPR	✓	-	-	-	-	-	✓	-
11	SOS [49] ◊	2015	CVPR	✓	✓	-	-	-	✓	-	-
12	MSO [49] ◊	2015	CVPR	✓	-	-	-	-	✓	-	-
13	XPIE [50] ◊	2017	CVPR	✓	✓	-	-	-	-	✓	-
14	ILSO [51] ◊	2017	CVPR	-	-	-	-	-	-	✓	✓
15	JOT [52] ◊	2017	FCS	✓	✓	✓	-	-	-	✓	-
16	DUTS [41] ♦	2017	CVPR	✓	✓	-	-	-	-	✓	-
17	**SOC (OUR)** ♦	2022	–	✓	✓	✓	✓	✓	✓	✓	✓

such as RGB-D SOD [82], [83], Event-RGB SOD (ERSOD) [4], Light Field SOD [84], Co-SOD [85], 360°Video SOD [86], and Video SOD [16].

2.2 SOD Datasets

In this section, we briefly discuss existing datasets designed for SOD tasks, focusing in particular on aspects including annotation type, number of salient objects per image, number of images, and image quality. These datasets are listed in Table 1.

Early datasets are either limited in their numbers of images or in their coarse annotations of salient objects. For example, salient objects in the original version of **MSRA-A** [32] and **MSRA-B** [32] are only roughly annotated in the form of bounding boxes. **ASD** [47], **SED1** [33] and **M10K** [35] contain only one salient object in most images, while the **SED2** [33] dataset provides two objects per image but contains only 100 images. In order to improve the quality of datasets, researchers in recent years have started to collect images with multiple objects in relatively complex and cluttered backgrounds. The new datasets include **ECC** [37], **DU-O** [38], **Judd-A** [36], and **PASCAL-S** [39]. These datasets are improved in terms of both annotation quality and number of images, compared to their predecessors. To resolve the shortcomings still present, some datasets (*e.g.*, **HKU** [40], **XPIE** [50], and **DUTS** [41]) provide large amounts of pixel-wise labeled images (Fig. 3.b) with more than one salient object per image. However, they ignore non-salient objects (1^{st} row in Fig. 1) and do not offer instance-level annotations (Fig. 3.c). Jiang *et al.* [52] collected roughly 6K *simple background images* (most of them are pure texture images) to cover non-salient scenes. However, their dataset, named **JOT**, falls short in capturing the complexity of real-world scenes. The dataset of **ILSO** [51] contains instance-level salient object annotations but only roughly labeled boundaries, as shown in Fig. 7. Beyond the "standard" SOD datasets, there are also several other special datasets that introduce new tasks, such as salient object subitizing (*i.e.*, **SOS** [49] and its subset **MSO** [49]).

To sum up, as discussed above, existing datasets mostly focus on images with clear salient objects and simple backgrounds. Considering the aforementioned limitations of existing datasets, a more realistic dataset, containing non-salient objects, textures "in the wild", and salient objects with attributes, is needed for future investigations in this field. Such a dataset could offer deeper insight into the strengths and weaknesses of SOD models, and help overcome performance saturation. Our **SOC** is unique in that it provides various high-quality annotations, as shown in Table 1.

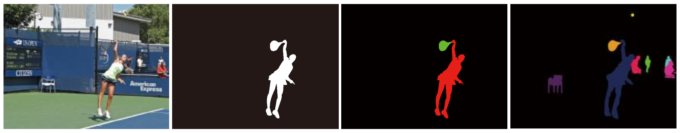

(a) Image (b) Previous (c) Ours (d) Segmentation

Figure 3. Previous SOD datasets only annotate the images by drawing pixel-accurate silhouettes around salient objects (b). Different from object segmentation datasets [87] (d) where (objects are not necessarily **salient**), our SOC provides *salient instances* (c). We provide a high-quality and large-scale annotated dataset comprised of images that better capture the properties of real-world scenes.

2.3 SOD Models

We have noticed that, from 1998 to the end of Feb. 2021, more than 10,000 papers on saliency detection or related field have been published. In this section, we try our best to summarize those published in top conferences (*e.g.*, NeurIPS, CVPR) and journals (*e.g.*, TPAMI, TIP), as well as some high-quality open-access (*i.e.*, arXiv) works. Instead of briefly describing the pipeline of each model, we summarize key components to provide a global view.

As shown in Table 2, a number of different approaches have been designed to tackle SOD using super-pixel, proposal, or edge/boundary annotations under different levels of supervision, such as unsupervised, semi-supervised, and fully supervised. Using common aggregation strategies (*e.g.*, linear, non-linear), these methods mainly focus on pixels, regions, and patches to design more powerful models. Besides, we note that certain priors (*e.g.*, the center-surround prior, local/global contrast prior, fore/background prior, and boundary prior) are frequently used in these methods. Some models also utilize different post-processing steps (*e.g.*, conditional random field, morphology, watershed, and max-flow strategies) to further improve the performance.

More recently, many deep learning SOD models based on different network architectures, such as multi-layer perceptrons, fully convolutional networks (FCNs), hybrid networks and capsules, have been proposed and achieve higher performance than traditional methods. According to the learning paradigm, most deep SOD models can be roughly split into two types: single-task learning and multi-task learning methods. We summarize the training data, backbones, and other components in Tables 3 and 4.

We mainly focus on macro-level statistics rather than micro-level descriptions. We kindly refer readers to the recent architecture review [46]. We hope this comprehensive review can serve as guidance[5] for future researchers in this fast-growing field.

[4] ERSOD: https://github.com/jxr326/ERSOD-Net.

[5] Research group: https://github.com/DengPingFan/Saliency-Authors.

Table 2
Summary of popular SOD models using handcrafted features. **Agg.:** Aggregation strategy, *e.g.*, LN = linear, NL = non-linear, HI = hierarchical, BA = Bayesian, AD = adaptive, LS = least-square solver, EM = energy minimization, and GMRF = Gaussian MRF. **SL.:** Supervision level, *e.g.*, unsupervised (★), semi-supervised (●), weakly supervised (☾), fully supervised (○), active learning (A). **Sp.:** Whether or not superpixel over-segmentation is used. **Pr.:** Whether or not proposal methods are used. **Ed.:** Whether or not edge cues are used. **Post-Pros.:** Whether post-processing methods (*e.g.*, CRF [88], graph-cut [89], GrabCut [90], Ncut [91]), morphology, max-flow (MF) [92] or only thresholding are used.

	#	Model	Publ.	Scholar	Prior.	Uniqueness	Component	Agg.	SL.	Sp.	Pr.	Ed.	Post-Pros
2010 - 1998	1	Itti [53]	TPAMI	link	center-surround	pixel	Color, Intensity, Orientation	LN	★	-	-	-	-
	2	GBVS [93]	NeurIPS	link	-	pixel	Markovian	-	★	-	-	-	-
	3	FT [47]	CVPR	link	frequency domain	pixel	Color, Luminance	-	★	-	-	-	-
	4	SR [94]	CVPR	link	spectral residual	pixel	Log Spectrum	-	★	-	-	-	-
	5	AIM [95]	NeurIPS	link	maximizing information	patch	Shannon's Self-information	-	★	-	-	-	-
	6	SUN [96]	JOV	link	self-information	pixel	DoG, ICA-derived features	-	★	-	-	-	-
	7	FG [97]	MM	link	local contrast	pixel	Fuzzy Growing	-	★	-	-	-	-
	8	AC [98]	ICVS	link	local contrast	multi-patch	Color, Luminance	LN	★	-	-	-	-
	9	SEG [99]	ECCV	link	local contrast	pixel	Conditional Probabilistic	-	★	-	-	-	CRF
	10	MSSS [100]	ICIP	link	symmetric surround	pixel	Color, Luminance	-	★	-	-	-	graph-cut
	11	ICC [101]	ICCV	link	isophote	global structure	curvedness, isocenters, color	LN	★	-	-	-	graph-cut
	12	EDS [102]	PR	link	-	pixel	threshold, distance, multi-DoG	-	★	-	-	✓	-
	13	RE [103]	ICME	link	local contrast	pixel/patch	Contrast pyramid	-	★	-	-	-	-
	14	RSA [104]	MM	link	global contrast	patch	Polar transfer, NN-GPCA [104]	-	★	-	-	-	-
	15	RU [105]	TMM	link	rule based	pixel	denoising, geometric	-	★	-	-	-	-
	16	CSM [106]	MM	link	frequency&contrast	pixel	Envelope, Skeleton	-	★	-	-	-	-
2014 - 2011	17	LSSC [107]	TIP	link	bayesian	pixel/region	convex hull, subspace clustering	NL	★	✓	-	-	-
	18	COV [108]	JOV	link	-	pixel/patch	covariance matrices	NL	★	-	-	-	-
	19	GR [109]	SPL	link	contrast, center, smooth	-	convex hull, continuous pair	NL	★	✓	-	-	-
	20	MSS [110]	SPL	link	local, integrity, center	-	various gaussian, convex hull	NL	★	✓	-	-	-
	21	LSMD [111]	AAAI	link	texture, edge, color	pixel/region	hierarchical clustering, gaussian	-	★	✓	✓	-	threshold
	22	BSF [112]	ICIP	link	boundary-based	region	convex hull, soft-segmentation	-	★	✓	-	-	-
	23	HC [113]	CVPR	link	global contrast	region	Histogram-based Contrast	-	★	-	-	-	graph-cut
	24	RC [113]	CVPR	link	global contrast	region	Region-based Contrast	-	★	-	-	-	graph-cut
	25	CA [113]	CVPR	link	context-aware	patch	Four principles	-	★	-	-	-	-
	26	MR [38]	CVPR	link	fore/back-ground	pixel/region	graph-based manifold ranking	-	★	✓	-	-	-
	27	SF [114]	CVPR	link	element contrast	region	uniqueness, spatial	NL	★	-	-	-	-
	28	HS [37]	CVPR	link	global contrast	hi-region	Region-scale, Location heuristic	HI	★	-	-	-	-
	29	DRFI [115]	CVPR	link	background descriptor	region	region vector, multi-level	LN	○	✓	-	-	-
	30	RBD [116]	CVPR	link	background weighted	region	background connectivity	LS	★	✓	-	-	-
	31	LR [117]	CVPR	link	location, semantic, color	pixel/region	Low rank matrix	NL	○	✓	-	-	threshold
	32	PCA [118]	CVPR	link	center-bias priors	patch	color, pattern, gaussian	NL	★	✓	-	-	-
	33	HDCT [119]	CVPR	link	high-dimensional color	pixel	Trimap, color transform	LN	★	✓	-	-	-
	34	CRFM [120]	CVPR	link	aggregation	pixel	GIST descriptor	NL	○	-	-	-	CRF
	35	STD [121]	CVPR	link	statistical textural	region	Graph, sparse texture	-	★	-	-	-	GrabCut
	36	PDE [122]	CVPR	link	representative elements	region	color, background, center	-	★	✓	-	-	-
	37	SUB [123]	CVPR	link	Submodular	region	color, spatial, center	-	○	✓	-	-	threshold
	38	PISA [124]	CVPR	link	spatial	pixel/region	color, structure, orientation	NL	★	-	-	✓	-
	39	DSR [125]	ICCV	link	reconstruction errors	multi-region	background, obj./centerGaussian	BA	★	✓	-	-	-
	40	MC [126]	ICCV	link	markov random walks	region	Markov Chain	-	★	✓	-	-	-
	41	GC [127]	ICCV	link	global cue	region	GMM, appearance, spatial	AD	★	-	-	-	-
	42	SVO [128]	ICCV	link	center-surround	patch/region	Graph, Obj.	EM	★	✓	✓	-	-
	43	CSD [129]	ICCV	link	center-surround	multi-patch	color, orientation, intensity	LN	★	-	-	-	-
	44	UFO [130]	ICCV	link	focus, objectness	pixel/region	Uniqueness, Focusness, Obj.	NL	★	✓	✓	✓	threshold
	45	CHM [131]	ICCV	link	center-surround, local	mRegion/patch	SVM, hyperedge	LN	●	✓	-	✓	threshold
	46	CIO [132]	ICCV	link	objectness	Region	Graph, frequency, Obj.	GMRF	★	✓	-	-	-
	47	CC [133]	ICCV	link	convexity context	mRegion	concavity, bounding box	-	★	✓	-	-	graph-cut
	48	GS [134]	ECCV	link	boundary, connectivity	patch/region	Geodesic distance transform	-	★	✓	-	✓	-
	49	CB [135]	BMVC	link	context, shape, center	mRegion	Iterative energy minimization	LN	★	✓	✓	-	-
	50	SLMR [136]	BMVC	link	low-rank matrix	Region	sparse noise	-	★	✓	-	-	-
2018 - 2015	51	SMD [137]	TPAMI	link	texture, edge, color	pixel/region	hierarchical clustering, gaussian	-	★	✓	✓	-	threshold
	52	RS [138]	TPAMI	link	fore/back-ground	region	manifold ranking, grouping cue	-	★	✓	-	-	-
	53	BFS [139]	NC	link	fore/back-ground seed	region	Gaussian falloff, threshold	NL	★	✓	-	-	-
	54	GLC [140]	PR	link	global/local contrast	region	HOG, LBP, codebook,graph-cut	LN	★	✓	-	-	-
	55	DSP [141]	PR	link	propagation	region	sink points, chi-square distance	NL	★	✓	-	✓	-
	56	LPS [142]	TIP	link	label propagation-base	pixel/region	three-cue-center, affinity matrix	NL	★	✓	-	-	-
	57	MAPM [143]	TIP	link	background	region	Markov absorption probability		★	✓	-	-	-
	58	MIL [144]	TIP	link	instance	region	multi-instance learning, SVM	-	●	✓	✓	-	-
	59	RCRR [145]	TIP	link	reversion correction	pixel/region	regular-random walks ranking	-	★	✓	-	-	-
	60	FCB [146]	TIP	link	fore/back-ground, center	region	color difference, color volume	NL	★	✓	-	-	-
	61	NCS [147]	TIP	link	center bias	pixel/region	Ncut, merging scheme	EM	★	✓	-	✓	Ncut
	62	MDC [148]	TIP	link	direction contrast	pixel	OTSU, morphological filter	NL	★	-	-	-	watershed
	63	HCCH [149]	TIP	link	closure completeness & reliability	object	hierarchical segmentation	NL	★	-	-	✓	-
	64	JLSE [150]	TIP	link	exemplar-aided	region	joint latent space embedding	-	○	✓	-	-	-
	65	IFC [151]	TMM	link	boundary homogeneity	pixel/region	linear feedback control system	-	★	✓	-	-	-
	66	NIO [152]	TNNLS	link	smoothness, boundary	region	graph, iterative optimization	BA	●	✓	-	-	-
	67	MBS [153]	ICCV	link	barrier distance	pixel	backgroundness cue	-	★	-	-	-	morphology
	68	GP [154]	ICCV	link	diffusion based	region/pixel	diffusion/laplacian matrix	-	★	✓	-	-	-
	69	BSCA [155]	CVPR	link	color/space contrast	region/pixel	cellular automata, bayesian	-	★	✓	-	-	OTSU [156]
	70	BL [157]	CVPR	link	image prior	mRegion	SVM, MKB [158], LBP	LN	○	✓	-	-	-
	71	MST [159]	CVPR	link	geometry information	pixel	minimum spanning tree	-	★	✓	-	-	morphology
	72	RRWR [160]	CVPR	link	error-boundary removal	pixel/region	regular-random walks ranking	-	★	✓	-	-	-
	73	TLLT [161]	CVPR	link	propagation,boundary	region	convex hull, teach-to-learn	-	★	✓	-	-	-
	74	WSC [162]	CVPR	link	weighted sparse coding	region	color histogram, dictionary	NL	★	✓	-	-	-
	75	PM [163]	ECCV	link	propagation	region	extended random walk	LN	★	✓	-	-	-
2021 - 2019	76	TSG [164]	TCSVT	link	regionally spatial consistency	region	Sparse Representation, graph	LN	★	✓	-	-	MF
	77	LFCS [61]	TCSVT	link	smoothness, boundary	region	Discrete Linear Control System	LN	●	✓	-	-	-
	78	AIGC [165]	TCSVT	link	contrast, object	region	irregular graph	-	★	✓	-	-	-
	79	FTOE [166]	TMM	link	contrast, center, distribute	pixel/region	fuzzy theory, object enhancement	LN	★	✓	✓	-	-
	80	MSGC [167]	TMM	link	fore/back-ground seed	region	multi-scale, global cue	NL	★	✓	-	-	-
	81	SIA [168]	TMM	link	boundary, dhs [169]	-	Cellular Automation	BA	★	✓	-	-	-
	82	KSR [170]	TIP	link	trained on [32]	region	R-CNN, Rank-SVM, subspace	-	A	-	✓	-	-
	83	MSR [171]	TIP	link	boundary connectivity	region	MBD [172]	-	★	✓	-	-	OTSU
	84	LRR [173]	TIP	link	background	pixel/region	Celluar Automata [155], FCN32	Metric	★	✓	-	-	-

2.4 Dataset-Enhancement Strategies for Deep Models

Existing deep SOD models focus on designing effective decoders [44], [59], [261], [262], [265], [279], [286] to aggregate features from different levels of the backbone network [198], [211], [293]. We argue that, as they employ a mapping function from the input training image set to the output training ground-truth set, deep models should also focus on dataset-enhancement strategies to improve model generalization ability. Three different strategies have been widely studied, including label smoothing [294], image augmentation [295], [296], and self-supervised learning [297].

Table 3
Summary of popular deep learning based SOD models. See Table 2 for more detailed descriptions. MB = MSRA-B dataset [32]. M10K = MSRA-10K [35] dataset. P-VOC2010 = PASCAL VOC 2010 semantic segmentation dataset [174]. CRF = Conditional random fields. **Clicking the scholar will link to the specific author's google scholar.**

	#	Model	Publ.	Scholar	#Training	Training Dataset	Backbone	SL.	Sp.	Pr.	Ed.	CRF
2015	1	SupCNN [175]	IJCV	link	800	ECC [37]	-	○	✓	-	-	-
	2	LEGS [176]	CVPR	link	340+3,000	PASCAL-S [39]+MB [32]	-	○	-	✓	-	-
	3	MDF [40]	CVPR	link	2,500	MB [32]	-	○	✓	-	✓	-
	4	MC [177]	CVPR	link	8,000	M10K [35]	GoogLeNet [178]	○	✓	-	-	-
2016	5	DSL [179]	TCSVT	link	(5,168+10,000)*80%	DU-O [38]+M10K [35]	LeNet [180]/VGGNet16	○	✓	-	-	-
	6	DISC [181]	TNNLS	link	9,000	M10K [35]	-	○	✓	-	-	-
	7	DS [182]	TIP	link	10,000	M10K [35]	VGGNet [183]	○	✓	-	✓	✓
	8	SSD [184]	ECCV	link	2,500	MB [32]	AlexNet [185]	○	✓	✓	-	-
	9	CRPSD [186]	ECCV	link	10,000	M10K [35]	VGGNet	○	✓	-	-	-
	10	RFCN [187]	ECCV	link	10,103+10,000	P-VOC2010 [174]+M10K [35]	VGGNet	○	✓	-	✓	-
	11	MAP [188]	CVPR	link	∼5,500	SOS [49]	VGGNet	○	-	✓	-	-
	12	SU [189]	CVPR	link	15,000+10,000	SALI [190]+M10K [35]	VGGNet	○	-	-	-	✓
	13	RACD [191]	CVPR	link	10,565	DU-O [38]+NJU [192]+NLP [193]	VGGNet	○	-	-	-	-
	14	ELD [194]	CVPR	link	9,000	M10K [35]	VGGNet	○	✓	-	-	-
	15	DHS [169]	CVPR	link	3,500+6,000	DU-O [38]+M10K [35]	VGGNet	○	-	-	-	-
	16	DCL [195]	CVPR	link	2,500	MB [32]	VGGNet	○	✓	-	-	✓
2017	17	DLS [196]	CVPR	link	10,000	M10K [35]	VGGNet	○	✓	-	-	-
	18	MSRNet [51]	CVPR	link	(500+)2,500+2,500	(ILSO [51]+)MB [32]+HKU [40]	VGGNet	○	-	✓	✓	✓
	19	SRM [197]	CVPR	link	10,553	DUTS [41]	ResNet50 [198]	○	-	-	-	-
	20	NLDF [199]	CVPR	link	2,500	MB [32]	VGGNet	○	-	-	✓	✓
	21	WSS [41]	CVPR	link	456K	ImageNet [200]	VGGNet	○	✓	-	✓	✓
	22	DSS [201]	CVPR	link	2,500	HKU [40]+MB [32]	VGGNet	○	-	-	✓	✓
	23	FSN [202]	ICCV	link	10,000	M10K [35]	VGGNet	○	-	-	-	-
	24	SVF [203]	ICCV	link	10,000	M10K [35]	VGGNet	☾	✓	-	-	-
	25	UCF [204]	ICCV	link	10,000	M10K [35]	VGGNet	○	-	-	-	-
	26	AMU [205]	ICCV	link	10,000	M10K [35]	VGGNet	○	-	-	✓	-
2018	27	EAR [206]	TCYB	link	2,500+2,500	HKU [40]+MB [32]	VGGNet16	○	-	-	-	-
	28	Refinet [207]	TMM	link	3,000	MB [32]	VGGNet16	○	✓	-	✓	✓
	29	LICNN [208]	AAAI	link	456K	ImageNet [200]	VGGNet	○	-	-	-	-
	30	ASMO [62]	AAAI	link	82,783+2,500+2,500	MsCO [87]+ HKU [40]+MB [32]	ResNet101	○	-	-	-	✓
	31	RADF [209]	AAAI	link	10,000	M10K [35]	VGGNet	○	-	-	-	✓
	32	R3Net [210]	IJCAI	link	10,000	M10K [35]	ResNeXt [211]	○	-	-	-	✓
	33	C2SNet [212]	ECCV	link	20,000+10,000	Web [212]+M10K [35]	VGGNet	○	✓	✓	-	-
	34	RAS [213]	ECCV	link	2,500	MB [32]	VGGNet	○	-	-	-	-
	35	LPSNet [214]	CVPR	link	10,553	DUTS [41]	VGGNet16	○	-	-	-	-
	36	RSOD [215]	CVPR	link	425	PASCAL-S [39]	ResNet101	○	-	✓	-	-
	37	DUS [66]	CVPR	link	2,500	MB [32]	ResNet101	☾	-	-	-	-
	38	ASNet [216]	CVPR	link	15,000+10,000+5,168	SALI [190]+M10K [35]+DU-O [38]	VGGNet	○	-	-	-	-
	39	BMPM [217]	CVPR	link	10,553	DUTS [41]	VGGNet	○	-	-	-	-
	40	DGRL [218]	CVPR	link	10,553	DUTS [41]	ResNet50	○	-	-	-	-
	41	PiCA [219]	CVPR	link	10,553	DUTS [41]	VGGNet16/ResNet50	○	-	-	-	✓
	42	PAGRN [220]	CVPR	link	10,553	DUTS [41]	VGGNet19	○	-	-	-	-
2019	43	SE2Net [221]	arXiv	link	10,553	DUTS [41]	VGGNet/ResNeXt101	○	-	-	-	-
	44	DRMC [222]	arXiv	link	10,533	DUTS [41]	VGGNet/ResNet101	○	-	-	-	✓
	45	RDSNet [223]	arXiv	link	10,000+10,553	M10K [35]+DUTS [41]	VGGNet/ResNet-152	○	-	-	-	✓
	46	AADF [224]	TCSVT	link	10,553	DUTS [41]	DenseNet161 [225]	○	-	-	-	-
	47	CCAL [226]	TMM	link	9,000	M10K [35]	VGGNet	○	-	-	-	-
	48	DeepUSPS [67]	NeurIPS	link	2,500	MB [32]	DRN-network [227]	☾	-	-	-	-
	49	FBG [228]	TIP	link	2,500	MB [32]	VGGNet16	○	-	-	✓	-
	50	SPA [229]	TIP	link	4,000	HKU [40]	-	○	✓	-	-	✓
	51	ConnNet [230]	TIP	link	2,500+2,500	MB [32]+ HKU [40]	ResNet50	○	-	-	-	-
	52	LFRWS [231]	TIP	link	10,000	M10K [35]	VGGNet16	○	-	-	✓	-
	53	RSR [72]	TPAMI	link	425	Extended of PASCAL-S [39]	ResNet101	○	-	-	-	-
	54	SSNet [232]	TPAMI	link	10,000	M10K [35]	VGGNet16	☾	✓	-	-	-
	55	LVNet [233]	TGRS	link	600	ORSSD [233]	-	○	-	-	-	-
	56	Deepside [234]	NC	link	2,500+10,553	MB [32]+DUTS [41]	VGGNet16	○	✓	-	-	-
	57	SuperVAE [235]	AAAI	link	-	-	VGGNet19	☾	✓	-	-	-
	58	DEF [236]	AAAI	link	10,553	DUTS [41]	ResNet101	○	-	-	-	-
	59	CapSal [63]	CVPR	link	82,783+5,265	MsCO [87]+COCO-CapSal [63]	ResNet101	☾	-	-	-	-
	60	MWS [237]	CVPR	link	300,000+10,553	ImageNet [200]+DUTS [41]	-	☾	✓	-	-	✓
	61	MLMS [238]	CVPR	link	10,553	DUTS [41]	VGGNet16	○	-	-	✓	-
	62	ICNet [239]	CVPR	link	10,000	M10K [35]	VGGNet16/ResNet50	○	-	-	-	✓
	63	AFNet [240]	CVPR	link	10,533	DUTS [41]	VGGNet16	○	-	-	✓	-
	64	PFANet [241]	CVPR	link	10,553	DUTS [41]	VGGNet16	○	-	-	✓	-
	65	PAGE [242]	CVPR	link	10,000	M10K [35]	VGGNet16	○	-	-	✓	✓
	66	CPD [243]	CVPR	link	10,533	DUTS [41]	VGGNet/ResNet50	○	-	-	-	-
	67	PoolNet [244]	CVPR	link	10,533	DUTS [41]	VGGNet/ResNet	○	-	-	✓	-
	68	BASNet [245]	CVPR	link	10,553	DUTS [41]	ResNet34/Xavier [246]	○	-	-	✓	-
	69	JDF [247]	ICCV	link	2,500	MB [32]	VGGNet16	○	-	-	✓	-
	70	DPOR [248]	ICCV	link	10,533	DUTS [41]	VGGNet16	○	-	-	-	-
	71	JLNet [249]	ICCV	link	10,582+10,533	P-VOC2010 [174]+DUTS [41]	DenseNet169	○	-	-	-	✓
	72	GLFN [58]	ICCV	link	1,600+10,533	HRSOD [58]+DUTS [41]	VGGNet	○	-	-	-	✓
	73	SIBA [250]	ICCV	link	10,533	DUTS [41]	ResNet50	○	-	-	✓	-
	74	SCRNet [44]	ICCV	link	10,533	DUTS [41]	ResNet50	○	-	-	✓	-
	75	EGNet [251]	ICCV	link	10,533	DUTS [41]	VGGNet/ResNet	○	-	-	✓	-
2020	76	HUAN [252]	TIP	link	10,553	DUTS [41]	VGGNet/ResNet/ResNetXt	○	-	-	-	✓
	77	ALM [253]	TIP	link	10,000+4,447	M10K [35]+ HKU [40]	DenseNet	○	✓	-	-	-
	78	HFFNet [254]	TIP	link	10,553	DUTS [41]	VGGNet16	○	-	-	✓	-
	79	DFI [255]	TIP	link	10,553	DUTS [41]	ResNet50	○	-	-	✓	-
	80	R2Net [256]	TIP	link	10,553	DUTS [41]	VGGNet16	○	-	-	-	-
	81	MRNet [257]	TIP	link	10,553	DUTS [41]	ResNet50	○	-	-	-	-
	82	CIG [258]	TIP	link	10,000	M10K [35]	VGGNet16	○	-	-	✓	-
	83	RASNet [259]	TIP	link	2,500	MB [32]	VGGNet16	○	-	-	-	-
	84	ASNet [260]	TPAMI	link	15,000+10,000+5,168	SALI [190]+M10K [35]+DU-O [38]	VGGNet	○	-	-	-	-
	85	DNNet [261]	TCYB	link	2,500+2,500	MB [32]+ HKU [40]	-	○	-	-	-	-
	86	CAANet [262]	TCYB	link	10,553	DUTS [41]	VGGNet16	○	-	-	-	-
	87	ROSA [263]	TCYB	link	2,500+5,168+2,500	HKU [40]+DU-O [38]+MB [32]	FCN [264]	○	✓	-	-	-
	88	DSRNet [265]	TCSVT	link	10,553	DUTS [41]	DenseNet	○	-	-	-	-
	89	EGNL [266]	TCSVT	link	2,500	MB [32]	VGGNet16	○	-	-	✓	-
	90	SACNet [267]	TCSVT	link	10,553	DUTS [41]	ResNet101	○	-	-	-	-
	91	FLGC [268]	TMM	link	10,553	DUTS [41]	VGGNet16	○	-	-	-	-
	92	TSNet [269]	TMM	link	4,000	MD4K [269]	ResNet50/VGGNet16	○	-	-	-	-

Instead of training directly with one-hot supervision, "label smoothing" techniques learn from smoothed supervision, and can thus relax the supervision signals using the generated smoothing labels [294] or disturbed labels [298]. Miyato *et al.* [299] applied local perturbations to data points to increase the smoothness of the model distribution. Thulasidasan *et al.* [300] discovered that mix-up training [296] with label smoothing can significantly improve model calibration. To obtain a more robust and generative

Table 4
Summary of popular deep learning based SOD models. See Tables 2 & 3 for more detailed descriptions.

	#	Model	Publ.	Scholar	#Training	Training Dataset	Backbone	SL.	Sp.	Pr.	Ed.	CRF
2020	93	SUCA [270]	TMM	link	10,553	DUTS [41]	ResNet50	○	-	-	-	-
	94	MIJR [271]	TMM	link	2,500+5,000	MB [32]+DUTS [41]	VGGNet16	○	✓	✓	-	✓
	95	CAGVgg [272]	PR	link	10,553	DUTS [41]	VGGNet/ResNet/NASNet [273]	○	-	-	-	-
	96	U2Net [274]	PR	link	10,553	DUTS [41]	UNet	○	-	-	-	-
	97	SalGAN [275]	TII	link	10,000	M10K [35]	VGGNet16	○	-	-	-	-
	98	ADA [276]	AAAI	link	2,500+780	MB [32]+NIR [276]	VGGNet16	○	-	-	-	-
	99	PFPNet [277]	AAAI	link	10,553	DUTS [41]	ResNet101	○	-	-	-	-
	100	GCPANet [278]	AAAI	link	10,553	DUTS [41]	ResNet50	○	-	-	-	-
	101	F3Net [279]	AAAI	link	10,553	DUTS [41]	ResNet50	○	-	-	✓	-
	102	LDF [280]	CVPR	link	10,553	DUTS [41]	ResNet50	○	-	-	✓	-
	103	ITSD [281]	CVPR	link	10,553	DUTS [41]	VGGNet16/ResNet50	○	-	-	✓	-
	104	SANet [64]	CVPR	link	10,553	DUTS [41]	VGGNet16	☾	-	-	✓	✓
	105	MINet [282]	CVPR	link	10,553	DUTS [41]	VGGNet16/ResNet50	○	-	-	-	-
	106	ABPNet [283]	ECCV	link	10,553	DUTS [41]	VGGNet16	○	-	-	✓	-
	107	CSNet [284]	ECCV	link	10,553	DUTS [41]	-	○	-	-	-	-
	108	GateNet [285]	ECCV	link	10,553	DUTS [41]	VGGNet16	○	-	-	-	✓
2021	109	DNA [286]	TCYB	link	10,553	DUTS [41]	VGGNet16/ResNet50	○	-	-	-	-
	110	DAFNet [59]	TIP	link	1,400	EORSSD [59]	VGGNet16	○	-	-	✓	-
	111	HGA [287]	TIP	link	10,553	DUTS [41]	VGGNet16	○	-	-	✓	-
	112	HIRN [288]	TIP	link	10,553	DUTS [41]	VGGNet16	○	-	-	✓	-
	113	SCWS [289]	AAAI	link	10,553	SDUTS [64]	ResNet50	☾	-	-	-	-
	114	PFS [290]	AAAI	link	10,553	DUTS [41]	ResNet50	○	-	-	✓	-
	115	KRNet [291]	AAAI	link	10,553	DUTS [41]	ResNet50	○	-	-	✓	-
	116	BAS [31]	arXiv	link	10,553	DUTS [41]	ResNet34	○	-	-	✓	-
	117	ICON [60]	arXiv	link	10,553	DUTS [41]	ResNet50	○	-	-	-	-
	118	ABP [292]	TPAMI	link	10,553	DUTS [41]	ResNet50	○	-	-	-	-
	119	CVAE [292]	TPAMI	link	10,553	DUTS [41]	ResNet50	○	-	-	-	-

model, Xie *et al.* [298] randomly replaced a portion of labels with incorrect values in each iteration. In addition, Wager *et al.* [301] demonstrated that corrupting training examples with noise from known distributions within the exponential family can inject appropriate generative assumptions into discriminative models, thus reducing generalization errors. Peterso *et al.* presented a soft-label dataset (CIFAR10H [302]) aiming at reflecting human perceptual uncertainty by providing label distributions across categories instead of hard one-hot labels.

Image augmentation [295] is an effective technique for extending the diversity of a training dataset, thus improving model generalization ability. Existing data augmentation techniques can be roughly divided into two categories: 1) human-designed policies, *e.g.*, rotation or scale transformation, and 2) learned policies [303], [304]. For the former, a predefined data augmentation policy is applied to the dataset. Beside the widely used rotation and scale transformations, other extensively studied methods in this category are erasing techniques [305], [306], which achieve data augmentation by randomly erasing part of the image patch. Further, mix-up methods [307], [308] utilize the mix-up data augmentation strategy to generate new samples from an existing training dataset to mitigate the uncertainty in prediction. For the latter [303], the network learns an image-conditioned data augmentation policy, which is usually parameterized by a deep neural network. In this way, the input image is fed to the data augmentation network to generate augmented samples with hyperparameters that control the degree of data augmentation.

Self-supervised learning [297], [309], also termed as consistency learning, defines an annotation-free pretext task to provide a surrogate supervision signal for feature learning. Conventionally, self-supervised learning is used for unsupervised representation learning to learn the feature embedding of the image or video. Recently, works have defined self-supervised learning as an auxiliary task, and used it within a weakly supervised [289] or semi-supervised learning framework [310]. Several recent and representative arts can be found in [311], [312], [313].

As far as we know, no existing salient object detection works have focused on exploring the dataset bias issue with dataset-enhancement strategies. In this paper, we claim that efforts on developing dataset-improvement strategies can also yield significant performance gains. Further, these solutions are general and can be easily applied to existing saliency detection networks.

3 SOC Dataset

In this section, we present details of our new challenging *SOC* dataset. Sample images from *SOC* are shown in Fig. 1, while statistics regarding the categories and attributes are shown in Fig. 4 (a) and Fig. 6, respectively. Based on the strengths and weaknesses of existing datasets, we identify seven crucial requirements that a comprehensive and balanced dataset should fulfill.

1) Presence of Non-Salient Objects. Most existing SOD datasets assume that an image should contain at least one salient object and thus discard images without salient objects [52]. However, this assumption is only true under ideal settings, which leads to *data selection bias*. In realistic settings, images do not always contain salient objects. For example, some images of amorphous backgrounds, such as sky, grass or textures, contain no salient objects at all [314]. The non-salient objects or background "stuff" may occupy the entire scene, and hence heavily constrain the possible locations of a salient object. Xia *et al.* [50] proposed a state-of-the-art SOD model the determines what is or is not a salient object, indicating that non-salient objects are crucial for reasoning salient objects. This suggests that non-salient objects deserve equal attention in SOD. Incorporating images containing non-salient objects makes a dataset more realistic, and hence more challenging. We define "*non-salient objects*" as images without salient objects or images with "stuff" categories. As suggested in [50], [314], the "stuff" categories include (a) densely distributed similar objects, (b) fuzzy shapes, and (c) regions without semantics, as illustrated in Fig. 5 (a)-(c), respectively.

To avoid data selection bias, we selected images randomly and automatically, as suggested by Torralba and Efros [43]. Based on the characteristics of non-salient objects, we randomly collected 783 texture images from the DTD [315] dataset. To enrich the diversity, 2,217 images including aurora, sky, crowds, store and many other kinds of realistic scenes were gathered from the Internet and other datasets [34], [39], [52], [87].

2) Number and Categories of Images. Providing a large number of images is essential for capturing the diversity and abundance of real-world scenes. Moreover, with large amounts of data, SOD models can avoid overfitting and enhance generalization.

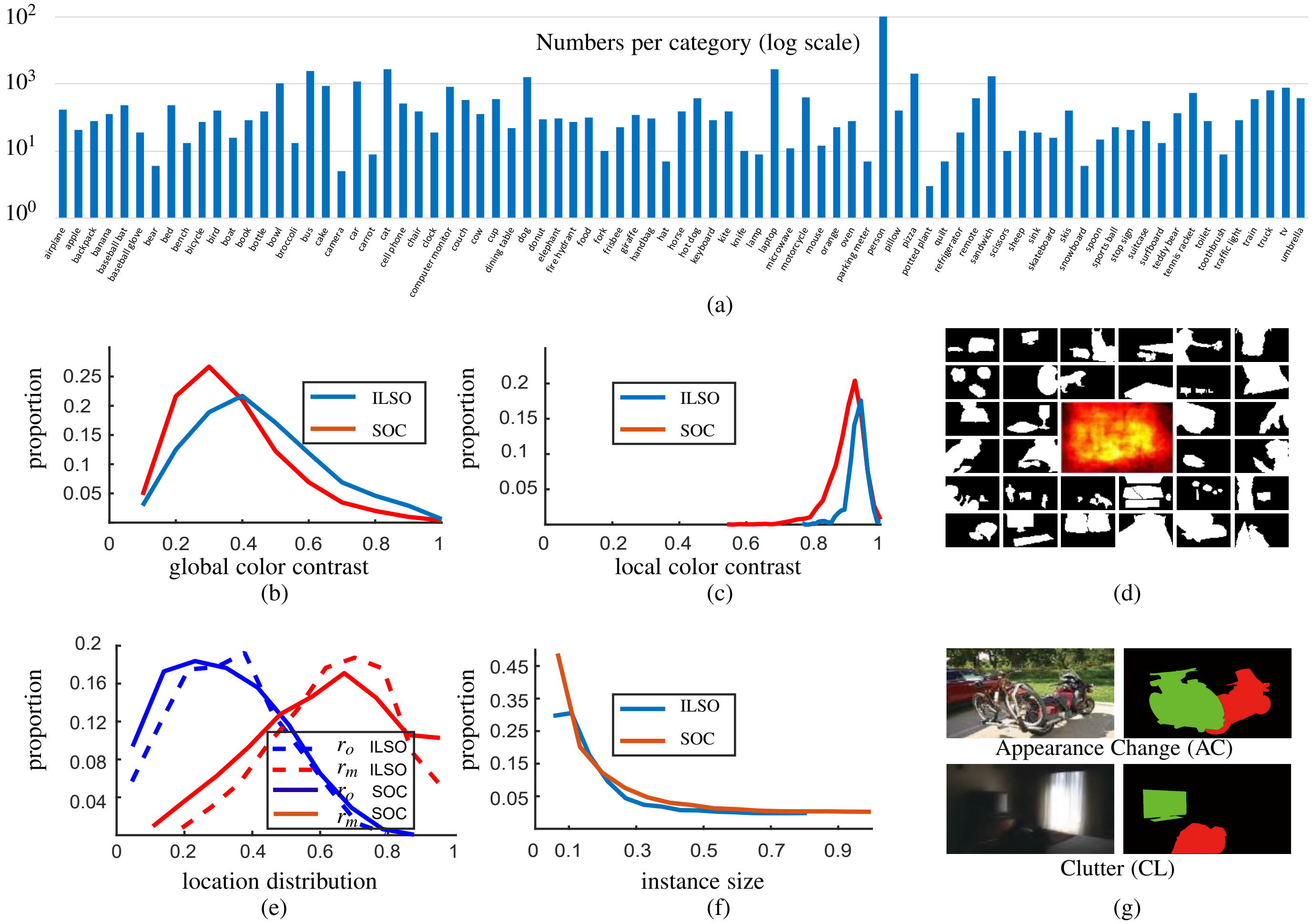


Figure 4. (a) Number of annotated instances per category in our SOC dataset. (b, c) Global and local color contrast statistics, respectively. (d) A set of saliency maps from our dataset and their overlay map. (e) Location distribution of the salient objects in SOC. (f) Distribution of instance sizes in the SOC and ILSO [51] datasets. (g) Visual examples of attributes. Best view on screen and zoomed-in for details.

To this end, we first randomly gathered 3,000 images from the MS-COCO dataset [87], which contains 'everyday scenes of common objects in their **natural** context.' Then, more than 80 object categories (see supplementary materials) were annotated. Note that we separated the process of data selection and labeling to avoid data selection bias, as discussed in [43]. Please refer to the subsection "*7) High-Quality Salient Object Labeling*" for details on this. Fig. 4 (a) shows the number of salient objects in each category. As can be seen, the "person" category accounts for a large proportion of the data, which is reasonable as people usually appear in daily scenes along with other objects. We divided our dataset (3k non-salient images and 3k salient images) into training, validation and test sets in the ratio of 6:2:2.

3) Global vs. Local Color Contrast of Salient Objects. As described in [39], the term "salient" is related to the global/local contrast of the foreground and background. It is essential to confirm whether the salient objects are easy to detect. For each object, we compute separate RGB color histograms for the foreground and background. Then, χ^2 distance is utilized to measure the distance between the two histograms. The global and local color contrast distributions are shown in Fig. 4 (b) and (c), respectively. Compared to ILSO, the SOC dataset has a higher proportion of objects with low global and local color contrast.

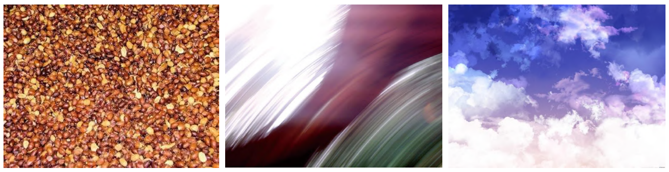
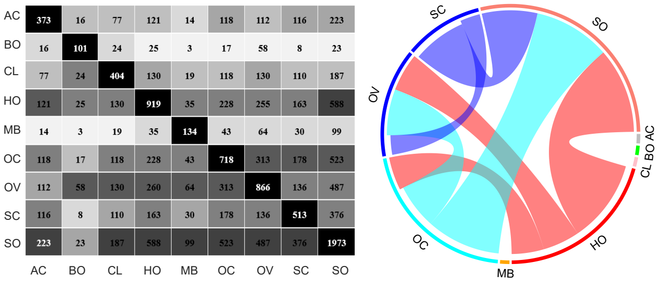


Figure 5. Examples of non-salient objects in our dataset. a) Crowded scene, b) motion blur, and c) background with non-interesting regions.

4) Locations. *Center bias* has been identified as one of the most significant and challenging biases pertaining to saliency detection datasets [39], [75], [316]. Fig. 4 (d) illustrates a set of images and their overlay map (*i.e.*, average mask map). As can be seen, although salient objects are located at different positions, the overlay map shows that somehow these images are still center biased. Unfortunately, previous benchmarks have often adopted this incorrect approach to analyze the positional distribution of salient objects. To avoid this misleading phenomenon, in Fig. 4 (e), we plot the statistics of two quantities, r_o and r_m, which denote how far an object center and its farthest (margin) point are from the image center, respectively. Both r_o and r_m are divided by half the diagonal length of the image for normalization, such that $r_o, r_m \in [0,1]$. From these statistics, we observe that the salient objects in our dataset do not suffer from center bias.

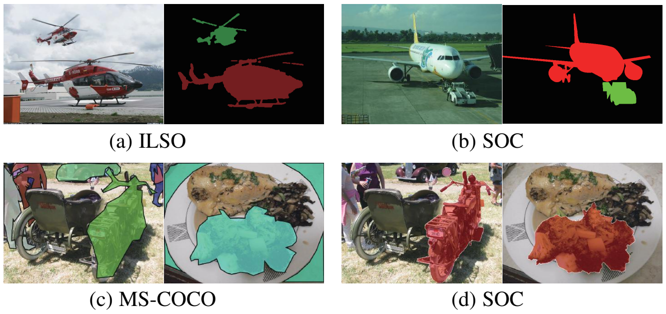
Table 5
List of salient object image attributes and their corresponding descriptions. These attributes are derived by studying the characteristics of existing datasets. Some visual examples can be found in Fig. 1 and Fig. 4 (g). For more examples, please refer to the supplementary materials.

Attribute	Description
AC ***(Appearance Change)***	Obvious illumination change in the object region.
BO ***(Big Object)***	The ratio between the object area and the image area is larger than 0.5.
CL ***(Clutter)***	Foreground and background regions around the object have similar colors. We labeled images with a global color contrast value larger than 0.2 and local color contrast value smaller than 0.9 as cluttered images (§ 3).
HO ***(Heterogeneous Objects)***	Objects composed of visually distinctive/dissimilar parts.
MB ***(Motion Blur)***	Objects have fuzzy boundaries due to camera shaking or motion.
OC ***(Occlusion)***	Objects are partially or fully occluded.
OV ***(Out-of-View)***	Part of the object is clipped by the image boundaries.
SC ***(Shape Complexity)***	Objects have complex boundaries, such as thin parts (*e.g.*, the foot of animal) and holes.
SO ***(Small Object)***	The ratio between the object area and the image area is smaller than 0.1.



	AC	BO	CL	HO	MB	OC	OV	SC	SO
AC	373	16	77	121	14	118	112	116	223
BO	16	101	24	25	3	17	58	8	23
CL	77	24	404	130	19	118	130	110	187
HO	121	25	130	919	35	228	255	163	588
MB	14	3	19	35	134	43	64	30	99
OC	118	17	118	228	43	718	313	178	523
OV	112	58	130	260	64	313	866	136	487
SC	116	8	110	163	30	178	136	513	376
SO	223	23	187	588	99	523	487	376	1973

Figure 6. Left: Attribute distribution over salient object images in our SOC dataset. Each number in the grid indicates the number of occurrences. Right: The dominant dependencies among attributes based on the frequency of occurrences. A larger link width indicates a higher probability of an attribute occurring with other ones.



(a) ILSO (b) SOC

(c) MS-COCO (d) SOC

Figure 7. Compared with the recent instance-level ILSO dataset [51] (a), which is labeled with discontinuous coarse boundaries, and MS-COCO dataset [87] (c), which is labeled with polygons, our SOC dataset (b & d) is labeled with smooth fine boundaries.

5) Size of Salient Objects. The size of an instance-level salient object is defined as the proportion of its pixels to those in the overall image [39]. As shown in Fig. 4 (f), the sizes of salient objects in our SOC vary greatly compared with the only other existing instance-level dataset, ILSO [51]. Further, there is a higher proportion of medium-sized objects in SOC.

6) Salient Objects with Attributes. Having attribute information for the images in a dataset helps objectively assess the performance of models over different types of parameters and variations. It also allows for the inspection of model failures. To this end, we define a set of attributes to represent specific situations encountered in common scenes, such as *motion blur*, *occlusion* and *cluttered background* (summarized in Table 5). Note that an image can be annotated with multiple attributes as these attributes are not exclusive.

Inspired by [317], we present the distribution of attributes over the dataset in the left of Fig. 6. The *SO* attribute makes up the largest proportion due to our accurate instance-level annotations (*e.g.*, the tennis racket in Fig. 3). The *HO* attribute also accounts for a large proportion, because the real-world scenes are composed of different constituent materials. *Motion blur (MB)* is more common in video frames, but also sometimes occurs in still images. Thus, *MB* images make up a relatively small proportion of our dataset. Since a realistic image usually contains multiple attributes, we show the dominant dependencies among attributes based on the frequency of occurrence on the right of Fig. 6. For example, a scene containing several heterogeneous objects is likely to have a large number of objects occluding each other and forming complex spatial structures. Thus, the *HO* attribute has a strong dependency with *OC*, *OV*, and *SO*.

7) High-Quality Salient Object Labeling. As noted in [318], training on the ECC dataset (1,000 images) yields better results than when using other datasets (*e.g.*, M10K with 10,000 images). This is because, besides the scale, dataset quality is also an important. To obtain a large number of high-quality images, we randomly selected images from the MS-COCO dataset [87], which is a large-scale challenging dataset whose objects are annotated with polygons (*i.e.*, coarse labeling). High-quality labels also play a critical role in improving the accuracy of SOD models [47]. Towards this end, we re-labeled the dataset with pixel-wise annotations. Following other famous task-oriented SOD benchmark datasets [32], [33], [34], [35], [37], [40], [41], [47], [50], [51], [52], we did not use an eye tracking device. We took two steps to ensure high-quality annotations: (i) We asked five viewers to annotate objects that they thought were salient in each image with bounding boxes (bboxes), and (ii) we kept the images in which the majority (≥ 3) of viewers annotated the same objects (IOU of the bbox > 0.8). After this first stage, we had 3,000 salient object images annotated with bboxes. *In the second stage*, we further manually labeled accurate silhouettes of the salient objects according to the bboxes. Note that we had 10 volunteers involved in both steps to cross-check the quality of annotations. In the end, we kept 3,000 images with high-quality, instance-level labeled salient objects. As shown in Fig. 7 (b & d), the boundaries of our object labels are precise, sharp and smooth. During the annotation process, we also added some new categories (*e.g.*, *computer monitor, hat, pillow*) that are not labeled in the MS-COCO dataset [87].

4 Our Dataset-Enhancement Strategies

Instead of focusing on designing a strong decoder for feature aggregation, we introduce three simple dataset-enhancement strategies to achieve better model generalization ability. We argue that the proposed strategies are easy to implement by existing fully supervised SOD models, and yield good performance with little effort. Let us define the RGB saliency training dataset as $D = \{x_i, y_i\}_{i=1}^{N}$, where x_i, y_i are an input RGB image and its corresponding ground-truth (GT) saliency map, i indexes the training images, and N is the size of the training dataset. As SOD is a binary prediction task, the GT saliency map y is usually a binary map, and most existing SOD techniques employ a binary (or weighted) cross-entropy loss function to evaluate the saliency prediction. In this paper, instead of defining the GT saliency map as a binary segmentation map, we first introduce "label smoothing" [294] as an effective technique to achieve both efficient model training and high model performance. Then, we adopt random image augmentation to generate diverse samples for better model generalization ability. Finally, as a widely studied technique in semi-supervised or unsupervised learning [297], [309], we extend the self-supervised learning solution to fully supervised SOD to achieve a robust model.

4.1 Label Smoothing

Label Smoothing and Knowledge Distillation. One of the most important scenarios in which to apply label smoothing is the teacher-student net [319] for knowledge distillation. Typically, in a teacher-student net, the teacher model has a strong learning capacity, while the student model has a lower one. The teacher model then teaches the student model by providing the latter with a "soft target". As discussed in [320], the "soft target" contains a rich similarity structure over the data, which is essential for producing an enhanced student model. Further, label smoothing can be treated as a form of output distribution regularization that prevents the network from overfitting. As pointed out in [294], hard labels may lead to the overfitting as the model will assign full probability to each category, which is not guaranteed to generalize well. With soft labels, the model learns the structure of the data, thus preventing it from being over-confident. Following the same data setting, *e.g.*, employing label smoothing, [321] introduced online label smoothing solution to gradually update the soft labels based on the model's prediction.

Conventional Setting. Given an input image x and the corresponding ground-truth saliency map y, the conventional deep saliency model f_θ is trained to achieve saliency prediction $s = f_\theta(x)$ by minimizing the cross-entropy loss: $\mathscr{L}_{ce}(y,s) = -\sum_{i=1}^{N}\sum_{u,v} y_i^{u,v} \log s_i^{u,v}$, where (u,v) index pixels. For the hard label based framework, we have $y \in \{0,1\}$, where 1 indicates the salient foreground and 0 represents the background.

Label Smoothing Setting. Different from the above hard label setting, in label smoothing regularization (LSR) [294], a smoothed label y' is used instead of y, which is formulated as:

$$y' = (1-\varepsilon)y + \varepsilon u(x). \tag{1}$$

Here, ε is the smoothing parameter, and $u(x)$ is a fixed distribution, which is usually defined as a uniform distribution. The smoothed label with a uniform distribution $u(x)$ is then defined as:

$$y' = (1-\varepsilon)y + \frac{\varepsilon}{K}, \tag{2}$$

where K is the number of categories.

Loss Function. Given smoothed label y' and hard label y, the loss function with LSR is defined as:

$$\mathscr{L}_{ls} = (1-\alpha)\mathscr{L}_{ce}(y,s) + \alpha\mathscr{L}_{ce}(y',s), \tag{3}$$

where α is used to balance the contribution of the smoothed and hard labels, and the smoothed label related loss is defined as $\mathscr{L}_{lsr} = \mathscr{L}_{ce}(y',s)$. Note that, if there exist other loss functions, the smoothed label can only be used in cross-entropy loss.

What Does Label Smoothing Really Do? The conventional cross-entropy loss can be rewritten as:

$$\mathscr{L}_{ce} = -\log s. \tag{4}$$

Here, s is the model prediction after sigmoid activation (for binary classification), which is defined as:

$$s_j = e^{z_j} / \sum_{k=1}^{K} e^{z_k} = 1/(1 + \sum_{k \neq j} e^{z_k - z_j}). \tag{5}$$

We then substitute s in (Eq. 4) and obtain:

$$\mathscr{L}_{ce} = \log(1 + \sum_{k \neq j} e^{z_k - z_j}). \tag{6}$$

Let us define the gap between the correct class and others as $M = z_k - z_j$. We can then conclude that the conventional cross-entropy loss aims to maximize this gap.

For label smoothing setting, as in (Eq. 2), we rewrite the smoothed label related loss $\mathscr{L}_{lsr}$ as:

$$\begin{aligned}\mathscr{L}_{lsr} &= -((1-\varepsilon)y + \varepsilon/K)\log s - (1-(1-\varepsilon)y - \varepsilon/K)\log(1-s) \\ &= -(y\log s + (1-y)\log(1-s)) + (\varepsilon y - \frac{\varepsilon}{K})\log(\frac{s}{1-s}).\end{aligned} \tag{7}$$

Using the definition of s in (Eq. 5), we have:

$$\frac{s_j}{1-s_j} = \frac{1}{\sum_{k=1}^{K} e^{z_k - z_j} - 1}. \tag{8}$$

We can then combine (Eq. 8) with (Eq. 7) and obtain:

$$\mathscr{L}_{lsr} = \mathscr{L}_{ce}(y,s) + (\varepsilon y - \frac{\varepsilon}{K}) * \frac{1}{\sum_{k=1}^{K} e^{z_k - z_j} - 1}. \tag{9}$$

The first part of (Eq. 9) aims to maximize the gap between the correct class and the others, which is same as the conventional binary-cross entropy loss as in (Eq. 6). The second part works in the opposite direction (compared with (Eq. 6)) to narrow the gap. In this way, the smoothed label related loss works to balance the gap between the correct class and others, which serves as an regularization to prevent the model from being over-confident.

4.2 Data Augmentation

As an effective data pre-processing technique, data augmentation aims to generate new samples from an existing dataset, thus producing a model with good generalization ability. Given the training dataset $D = \{x_i, y_i\}_{i=1}^{N}$, data augmentation produces a new dataset $D' = \{x_i', y_i'\}_{i=1}^{N'}$. As discussed previously, two main types of data augmentation have received particular attention. These include the handcrafted policies and learned policies [303], [304]. For the learned policies, we observe that the augmented data can change drastically depending on the context, which may not be an issue for image classification, but will change the salient attributes of an image. We thus focus only on handcrafted policies.

For handcrafted data augmentation policies, existing works [305], [306], [307], [308] focus on three main directions: 1) image transformation, *e.g.*, scale or rotation transformation; 2) mix-up to generate new samples, which are neighbors of the existing samples; and 3) adding noise to the ground-truth. Similar to learned policies, the mix-up strategy change the context information of an image, which is harmful for context-based tasks, such as salient object detection. In this paper, we therefore focus on two very simple data augmentation techniques, namely image transformation and adding noise to the ground-truth. For image transformation, we randomly scale, rotate and crop part of the image (85% of the original image to keep the context information). For the additive noise solution, we randomly add Gaussian noise of distribution $\mathcal{N}(0.1, 0.3)$ to the ground-truth saliency map, leading to a noisy ground-truth map. Note that, for image transformation, we transform image and ground-truth at the same time, while when adding noise to the ground-truth, we only process the ground-truth saliency maps.

4.3 Self-Supervised Learning

Self-supervised learning learns from an image without knowing the task itself or the ground-truth, making it an unsupervised representation learning technique. Conventionally, for the supervised learning setting, the loss function is defined as $\mathscr{L}_{ce}(y,s)$, where s is the model's prediction, and y is the ground-truth map. For self-supervised learning, the final loss function usually includes two main parts: the conventional cross-entropy loss $\mathscr{L}_{ce}(y,s)$ and an unsupervised loss that serves as a regularizer, *i.e.*, $\mathscr{L}(g(x),s)$, where $g(x)$ is the transformation of the original input image x. The two studies [297], [310] introduced a self-supervised loss with rotation estimation as a pretext task.

Similarly, we introduce a scale/rotation consistency loss function to achieve scale/rotation invariant predictions. Specifically, given an input image x, we define its prediction as s. Then, we apply an image transformation (scale or rotation transformation) and obtain x^t. We then perform the same transformation on the prediction s and obtain s'. We feed x^t to the same salient object detection network to get the saliency prediction as s^t. We assume that s' and s^t should be similar. Then, we adopt the single scale structural similarity index measure (SSIM) [322], [323] as a similarity measure, and define the self-supervised loss as:

$$\mathscr{L}_{ss} = 1 - SSIM(s', s^t). \tag{10}$$

4.4 Loss Function with the Proposed Strategies.

With the three introduced data-enhancement strategies, we first apply random data augmentation to both our training image set and training ground-truth set, as in Section 4.2. Then we generate the smoothed label following (Eq. 1), with $K = 2$ in this paper to represent the salient foreground and background regions. In addition to the loss function in (Eq. 3), we also introduce a self-supervised loss $\mathscr{L}_{ss}$. Our final loss function is then defined as:

$$\mathscr{L} = \mathscr{L}_{ls} + \gamma \mathscr{L}_{ss}, \tag{11}$$

where γ is introduced to balance the self-supervised loss, and is empirically set to $\gamma = 0.3$ in this paper.

5 SOC Benchmark

Based on three criteria (*i.e.*, representative pipeline, open-sourced, and state-of-the-art performance), we select 46 traditional SOD methods and 54 deep learning models from 203 reviewed methods (see § 2) to conduct our benchmark. To the best of our knowledge, this benchmark is the most comprehensive study in the RGB SOD.

5.1 Experimental Setup

5.1.1 Evaluation Metrics

Note that the GTs of non-salient images in our SOC dataset are all-zero matrices, so directly using the traditional F-measure [47] will result in very low and inaccurate scores. Thus, we utilize three golden metrics (*i.e.*, MAE [324], maximum E-measure [3], and S-measure [2]) to avoid this issue and to provide a more reliable assessment. Evaluation toolboxes are now publicly available.[6]

- **MAE** (M) is the mean absolute error metric, which is widely used to measure the pixel-level difference between the prediction and the GT.
- **E-measure** (E_{ξ}^{max}) is a new perceptual metric that takes both local and global similarity into consideration.
- **S-measure** (S_{α}) is a standard metric that quantizes the structural similarity at a region and object level.

Table 6
SOC dataset used in the benchmarking experiments.

	SOC_train	SOC_val	SOC_test	Total
Salient Objects (Sal)	1,800	600	600	3,000
Non-Salient Objects (NonSal)	1,800	600	600	3,000
Total	3,600	1,200	1,200	6,000

5.1.2 Training and Testing Protocols

The statistics of the SOC dataset used in the benchmark are summarized in Table 6. For traditional algorithms, we directly test their performance on the SOC-test set (1,200 images). For deep learning models, we first adopt the pre-trained models with their recommended training parameter settings under the default training dataset (see Tables 3 & 4) and then evaluate them on the SOC_test set to roughly obtain the 100 representative models (see Table 7 & 8). Finally, we provide a quantitative comparison and detailed analysis of 15 SOTA approaches, including the top-5 traditional methods and top-10 deep learning models.

5.2 Quantitative Comparisons

To build a standardized leaderboard (*i.e.*, same image resolution, thresholding step, and evaluation tool), we provide three golden metrics, *i.e.*, S_{α}, E_{ξ}^{max}, and M.

Table 7 shows the performance of 46 SOTA traditional SOD algorithms on our SOC_test set. In terms of both **S-measure** (*i.e.*, S_{α}) and max **E-measure** (E_{ξ}^{max}), the HCCH method surpasses all competitors by a large margin. RBD, COV, and DRFI obtain comparable performance in terms of S_{α} score. Meanwhile, COV ranks third in terms of S_{α} measure, but ninth in E_{ξ}^{max}. In terms of **MAE** (*i.e.*, M), the top-5 approaches are: SF, COV, HCCH, SR, and MSSS. It is worth mentioning that SF reduces M and outperforms all the recent traditional SOD methods. Based on their overall scores, the top-5 methods are HCCH, RBD, COV, DRFI, and WSC.

The quantitative results of the 54 deep learning SOD models on our SOC_test dataset are shown in Table 8. In terms of S_{α}, EGNet, R2Net, and CPDVgg are the top-3 models, with scores of more than 0.85. Roughly 46% (*i.e.*, 21/45) of model scores are between 0.650 and 0.800. Compared with the traditional model, which achieves an S_{α} score of 0.736, we can see continuous improvement over the past few years, with the exception of four early models (*i.e.*,

[6] https://github.com/mczhuge/SOCToolbox.

Table 7
Comparison of the traditional SOD algorithms on our SOC_test set (1,200 images) in terms of $S_\alpha \uparrow$, $E_\xi^{max} \uparrow$, and $M \downarrow$. The top-3 results are highlighted in **red**, **blue** and **green**, respectively. The superscript of each score is the corresponding ranking. Details of these methods are summarized in Table 2. The overall rank index indicates the average ranking of the three metrics. These results are available at: Google Drive.

	#	Model	Code	$S_\alpha \uparrow$	$E_\xi^{max} \uparrow$	$M \downarrow$	Rank
2014-before	1	SUN [96]	Matlab	0.475^{46}	0.688^{44}	0.436^{46}	46
	2	LSSC [107]	Matlab + C	0.552^{45}	0.714^{43}	0.365^{45}	45
	3	BSF [112]	Matlab	0.554^{44}	0.728^{38}	0.353^{44}	44
	4	GR [109]	Matlab + C	0.588^{41}	0.715^{42}	0.332^{42}	43
	5	HS [37]	EXE	0.601^{40}	0.729^{37}	0.321^{41}	42
	6	Itti [53]	Matlab	0.587^{42}	0.736^{30}	0.311^{39}	41
	7	AIM [95]	Matlab	0.605^{39}	0.670^{45}	0.250^{24}	39
	8	GBVS [93]	Matlab	0.615^{36}	0.733^{35}	0.293^{37}	39
	9	LR [117]	Matlab	0.642^{31}	0.723^{40}	0.253^{27}	36
	10	CA [325]	Matlab + C	0.606^{38}	0.750^{22}	0.291^{36}	35
	11	MR [38]	Matlab + C	0.645^{29}	0.734^{33}	0.259^{31}	32
	12	SEG [99]	Matlab + C	0.576^{43}	0.765^{7}	0.352^{43}	32
	13	FT [47]	C	0.626^{34}	0.738^{29}	0.236^{20}	28
	14	MC [126]	Matlab + C	0.656^{23}	0.736^{30}	0.251^{25}	26
	15	CB [135]	Matlab + C	0.653^{25}	0.758^{13}	0.268^{33}	23
	16	SR [94]	Matlab/C++	0.658^{21}	0.661^{46}	0.156^{4}	23
	17	PCA [118]	Matlab + C	0.670^{18}	0.741^{28}	0.209^{13}	17
	18	MSS [110]	Matlab	0.682^{12}	0.776^{4}	0.231^{19}	10
	19	SF [114]	C	0.699^{6}	0.747^{26}	$\mathbf{0.130}^{1}$	8
	20	DSR [125]	Matlab + C	0.702^{5}	0.751^{20}	0.184^{8}	8
	21	MSSS [100]	C	0.683^{11}	0.757^{14}	0.164^{5}	7
	22	HDCT [119]	Matlab	0.696^{7}	0.774^{5}	0.201^{12}	6
	23	DRFI [115]	C	0.709^{4}	$\mathbf{0.791}^{2}$	0.197^{11}	4
	24	COV [108]	Matlab	$\mathbf{0.711}^{3}$	0.761^{9}	$\mathbf{0.146}^{2}$	2
	25	RBD [116]	Matlab	$\mathbf{0.716}^{2}$	$\mathbf{0.784}^{3}$	0.186^{9}	2
2021-2015	26	WMR [326]	Matlab + C	0.640^{32}	0.733^{35}	0.269^{34}	38
	27	MAPM [143]	Matlab + C	0.644^{30}	0.722^{41}	0.256^{29}	37
	28	BL [157]	Matlab + C	0.623^{35}	0.751^{20}	0.296^{38}	32
	29	RRWR [160]	Matlab	0.647^{27}	0.735^{32}	0.258^{30}	31
	30	WLRR [327]	Matlab + C	0.614^{37}	0.759^{11}	0.312^{40}	30
	31	RCRR [145]	Matlab	0.650^{26}	0.734^{33}	0.255^{28}	29
	32	GP [154]	Matlab + C	0.632^{33}	0.759^{11}	0.287^{35}	27
	33	TLLT [161]	Matlab	0.656^{23}	0.725^{39}	0.214^{15}	25
	34	BSCA [155]	Matlab + C	0.657^{22}	0.755^{16}	0.259^{31}	22
	35	SMD [137]	Matlab	0.662^{20}	0.748^{25}	0.246^{22}	21
	36	MDC [148]	C	0.675^{16}	0.744^{27}	0.219^{17}	20
	37	DSP [141]	Matlab + C	0.664^{19}	0.754^{17}	0.248^{23}	17
	38	MIL [144]	Matlab + C	0.671^{17}	0.750^{22}	0.236^{20}	17
	39	MST [159]	C	0.647^{27}	0.773^{6}	0.251^{25}	16
	40	GLC [140]	Matlab + C	0.676^{15}	0.756^{15}	0.223^{18}	15
	41	MBS [153]	Matlab	0.678^{14}	0.753^{18}	0.214^{15}	14
	42	LPS [142]	Matlab + C	0.694^{9}	0.749^{24}	0.183^{7}	13
	43	WFD [328]	C	0.680^{13}	0.760^{10}	0.213^{14}	12
	44	BFS [139]	Matlab + C	0.696^{7}	0.753^{18}	0.195^{10}	10
	45	WSC [162]	Matlab	0.693^{10}	0.765^{7}	0.179^{6}	5
	46	HCCH [149]	Matlab	$\mathbf{0.736}^{1}$	$\mathbf{0.794}^{1}$	$\mathbf{0.149}^{3}$	1

Table 8
Evaluation of 54 deep learning based SOD models on our SOC_test set (1,200 images). We adopt the default implementations listed in Table 3 and Table 4 to test their generalization capability. These results are available at: Google Drive.

	#	Model	Code	$S_\alpha \uparrow$	$E_\xi^{max} \uparrow$	$M \downarrow$	Rank
2015	1	LEGS [176]	Caffe	0.679^{53}	0.765^{54}	0.228^{53}	54
	2	MDF [40]	Caffe	0.739^{49}	0.768^{53}	0.144^{43}	49
	3	MC [177]	Caffe	0.757^{47}	0.823^{43}	0.138^{35}	43
2016	4	DSL [179]	Caffe	0.724^{52}	0.810^{47}	0.194^{52}	51
	5	DISC [181]	Caffe	0.735^{51}	0.810^{47}	0.175^{50}	50
	6	DCL [195]	Caffe	0.771^{44}	0.836^{39}	0.157^{48}	45
	7	ELD [194]	Caffe	0.774^{42}	0.836^{39}	0.138^{35}	40
	8	DS [182]	Caffe	0.779^{40}	0.860^{24}	0.155^{46}	37
	9	DHS [169]	Pytorch	0.800^{32}	0.848^{33}	0.122^{30}	33
	10	RFCN [187]	Caffe	0.814^{23}	0.858^{27}	0.113^{23}	25
2017	11	UCF [204]	Caffe	0.654^{54}	0.805^{51}	0.285^{54}	53
	12	AMU [205]	Caffe	0.737^{50}	0.808^{50}	0.185^{51}	51
	13	SVF [203]	Caffe	0.761^{45}	0.816^{45}	0.156^{47}	47
	14	WSS [41]	Caffe	0.778^{41}	0.821^{44}	0.140^{39}	42
	15	DSS [201]	Caffe	0.807^{30}	0.858^{27}	0.111^{20}	27
	16	SRM [197]	Caffe	0.822^{16}	0.859^{26}	0.111^{20}	21
	17	MSRNet [51]	Caffe	0.816^{19}	0.871^{16}	0.117^{25}	20
	18	NLDF [199]	Tensorflow	0.816^{19}	0.860^{24}	0.104^{13}	16
2018	19	RAS [213]	Pytorch	0.759^{46}	0.813^{46}	0.151^{44}	46
	20	R3Net [210]	Pytorch	0.773^{43}	0.825^{42}	0.138^{35}	41
	21	LPSNet [214]	Pytorch	0.795^{35}	0.838^{38}	0.143^{42}	39
	22	DGRL-GLN [218]	Caffe	0.794^{36}	0.845^{36}	0.141^{40}	38
	23	C2SNet [212]	Caffe	0.791^{37}	0.845^{36}	0.138^{35}	36
	24	PiCA-Res [219]	Pytorch	0.810^{28}	0.858^{27}	0.128^{31}	31
	25	BMPM [217]	Tensorflow	0.810^{28}	0.853^{30}	0.119^{27}	29
	26	ASNet [216]	Keras	0.817^{18}	0.865^{20}	0.111^{20}	17
2019	27	MWS [237]	Pytorch	0.757^{47}	0.828^{41}	0.172^{49}	47
	28	AFNet [240]	Caffe	0.812^{24}	0.850^{32}	0.120^{29}	29
	29	SIBA [250]	Caffe	0.800^{32}	0.884^{10}	0.130^{33}	26
	30	Deepside [234]	Caffe	0.815^{21}	0.861^{23}	0.119^{27}	24
	31	PFANet [241]	Tensorflow	0.815^{21}	0.846^{35}	0.101^{8}	22
	32	PoolNet [244]	Pytorch	0.829^{13}	0.868^{18}	0.106^{16}	14
	33	SCRNet [44]	Pytorch	0.833^{11}	0.872^{15}	0.105^{14}	13
	34	CPDVgg [243]	Pytorch	$\mathbf{0.856}^{3}$	0.889^{6}	$\mathbf{0.079}^{2}$	2
	35	EGNet [251]	Pytorch	$\mathbf{0.858}^{1}$	$\mathbf{0.896}^{2}$	$\mathbf{0.078}^{1}$	1
2020	36	ABPNet [283]	Pytorch	0.783^{38}	0.810^{47}	0.153^{45}	44
	37	U2Net [274]	Pytorch	0.780^{39}	0.795^{52}	0.105^{14}	35
	38	GCPANet [278]	Pytorch	0.807^{30}	0.848^{33}	0.133^{34}	34
	39	ITSD [281]	Pytorch	0.798^{34}	0.870^{17}	0.142^{41}	32
	40	MINet [282]	Pytorch	0.819^{17}	0.864^{22}	0.117^{25}	22
	41	SANet [64]	Pytorch	0.812^{24}	0.868^{18}	0.106^{16}	17
	42	GateNetVgg [285]	Pytorch	0.827^{15}	0.865^{20}	0.108^{18}	15
	43	F3Net [279]	Pytorch	0.828^{14}	0.891^{5}	0.109^{19}	12
	44	CSNet [284]	Pytorch	0.834^{10}	0.876^{14}	0.103^{10}	11
	45	LDF [280]	Pytorch	0.835^{9}	0.878^{12}	0.103^{10}	10
	46	RASNet [259]	Pytorch	0.832^{12}	0.887^{8}	0.103^{10}	9
	47	CAGVgg [272]	Keras	0.837^{8}	0.878^{12}	0.088^{4}	8
	48	DFI [255]	Pytorch	0.838^{7}	$\mathbf{0.903}^{1}$	0.101^{8}	5
	49	R2Net [256]	Pytorch	$\mathbf{0.857}^{2}$	0.885^{9}	$\mathbf{0.084}^{3}$	4
2021	50	SCWS [289]	Pytorch	0.811^{26}	0.851^{31}	0.115^{24}	28
	51	ICON [60]	Pytorch	0.811^{26}	$\mathbf{0.896}^{2}$	0.128^{31}	19
	52	BAS [31]	Pytorch	0.842^{5}	0.882^{11}	0.092^{7}	7
	53	ABP [292]	Pytorch	0.842^{5}	0.889^{6}	0.091^{6}	6
	54	CVAE [292]	Pytorch	0.849^{4}	$\mathbf{0.892}^{4}$	0.089^{5}	3

DISC, DSL, LEGS, and UCF). At the same time, 30 out of 45 models achieve high performance (*e.g.*, 0.800≤ S_α ≤0.850) and the average performance is nearly 0.820. Interestingly, in terms of E_ξ^{max}, the multi-task learning framework DFI and integrity learning model have the best and second-best scores of 0.903 and 0.896, respectively. Consistent with S-measure, in terms of **MAE**, we obtain the same top-3 models EGNet, CPDVgg, and R2Net. From our 54 benchmarked models, we find that models that perform well in terms of S-measure also do well in MAE. Overall, the top-10 approaches are EGNet, CPDVgg, CVAE, R2Net, DFI, ABP, BAS, CAGVgg, RASNet, and LDF. In the following section (§ 6), we will provide a more detailed analysis of these models.

5.3 Qualitative Comparisons

Two qualitative comparisons are presented in Fig. 8 and Fig. 9. As can be seen from Fig. 8, deep models generate saliency maps that are similar to the GTs, to varying degrees. Specifically, for ASNet, C2SNet, BMPM, DCL, DHS, DSS, DS, DISC, SVF, RFCN, and PFANet, the position of the object can be well-identified. However, all these methods generate blurred responses on object boundaries. PFANet, MDF, MC and LEGS even nearly destroy the integrity of the object. To better highlight these results, we introduce yellow rectangles to mark the high-quality segmentation regions and utilize red arrows to point out the errors. We observe that eight models (ABPNet, AFNet, AMU, NLDF, RAS, SCWS, UCF, and WSS) can localize the human object but introduce additional noise. We also notice that CAGNet, CSNet, MINet, DGRL, EGNet, F3Net, ICON, PoolNet, and R3Net can even capture the small structure of the human elbow. Moreover, saliency maps from R2Net, Deepside, SIBA, and MSRNet demonstrate better results than the above-mentioned methods. Amazingly, BAS, U2Net, ABP, CPD, GateNet, GCPANet, ITSD, LDF, SCRN, and CAVE perform very close to the GT and result in knife-edge-shaped boundaries in the yellow rectangle region without any additional noise.

In sharp contrast to the deep learning models, the traditional models (Fig. 9) all fail without exception. WSC, HCCH, and RBD are the three most promising approaches. However, their results

Table 9
Comparison of 14 state-of-the-art approaches in terms of attribute-level performance. For deep learning models, we re-train them on our SOC-Sal_train set (*i.e.*, 1,800 images). Please refer to Tables 2, 3, & 4 for more details. These results are available at: Google Drive.

	Attribute / Model	AC		BO		CL		HO		MB		OC		OV		SC		SO		Avg.	
		$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$	$S_\alpha \uparrow$	$M \downarrow$
Traditional	COV [108]	0.505	0.216	0.277	0.577	0.453	0.280	0.508	0.229	0.494	0.219	0.484	0.246	0.423	0.314	0.535	0.174	0.525	0.172	0.467	0.270
	WSC [162]	0.541	0.205	0.356	0.517	0.517	0.252	0.556	0.211	0.536	0.210	0.529	0.227	0.475	0.292	0.567	0.170	0.535	0.181	0.512	0.252
	HCCH [149]	0.585	0.199	0.354	.0.525	0.537	0.254	0.615	0.197	0.547	0.202	0.552	0.225	0.468	0.298	0.595	0.165	0.588	0.162	0.538	0.247
	DRFI [115]	0.598	0.229	0.391	0.513	0.570	0.274	0.618	0.230	0.556	0.230	0.577	0.248	0.527	0.304	0.614	0.188	0.585	0.197	0.560	0.268
	RBD [116]	0.589	0.225	0.429	0.481	0.575	0.260	0.625	0.216	0.557	0.213	0.583	0.235	0.521	0.295	0.602	0.191	0.579	0.192	0.562	0.256
Deep Learning	ABP [292]	0.767	0.092	0.592	0.315	0.742	0.125	0.787	0.101	0.742	0.095	0.740	0.112	0.746	0.132	0.759	0.083	0.741	0.080	0.735	0.126
	EGNet [251]	0.791	0.088	0.593	0.307	0.739	0.137	0.788	0.110	0.763	0.115	0.743	0.120	0.750	0.138	0.800	0.076	0.753	0.088	0.747	0.131
	CPDVgg [243]	0.806	0.076	0.626	0.278	0.765	0.118	0.808	0.096	0.786	0.097	0.765	0.103	0.760	0.127	0.801	**0.070**	0.765	0.076	0.765	0.116
	CAGVgg [272]	0.795	0.080	**0.700**	**0.208**	0.782	0.115	0.808	0.098	0.764	0.102	0.751	0.120	0.763	0.127	0.795	0.081	0.744	0.093	0.767	0.114
	RASNet [259]	**0.821**	**0.066**	0.626	0.276	0.785	**0.106**	**0.816**	**0.087**	0.788	**0.086**	0.776	**0.096**	0.779	**0.113**	**0.810**	**0.066**	0.774	**0.070**	0.772	0.107
	CVAE [292]	0.813	0.075	0.688	0.217	**0.790**	**0.107**	**0.816**	**0.092**	0.784	0.091	0.771	0.104	0.776	0.115	**0.820**	**0.069**	0.767	0.080	**0.781**	**0.106**
	LDF [280]	0.819	**0.071**	**0.697**	**0.212**	**0.796**	**0.105**	**0.824**	**0.088**	**0.792**	**0.085**	**0.781**	**0.098**	**0.790**	**0.107**	0.780	0.073	**0.801**	**0.072**	**0.787**	**0.101**
	R2Net [256]	**0.827**	**0.071**	0.656	0.257	**0.802**	**0.107**	**0.826**	**0.092**	**0.794**	0.097	**0.789**	**0.099**	**0.791**	**0.112**	**0.807**	0.072	**0.788**	**0.073**	**0.787**	0.109
	BAS [31]	**0.831**	**0.060**	**0.723**	**0.166**	0.785	0.110	0.814	0.093	**0.797**	**0.072**	**0.780**	0.101	**0.781**	0.114	**0.820**	0.072	**0.787**	0.075	**0.791**	**0.096**
	Avg.	0.721	0.125	0.551	0.346	0.688	0.168	0.729	0.139	0.693	0.137	0.687	0.152	0.668	0.185	0.722	0.111	0.693	0.115	-	-

are still far from the GT map, since they are mainly based on various prior features extracted from color, orientation, contrast, *etc*. Further, the center bias prior is not suitable in this case, since the human is located close to the image boundary, thus making this example more challenging for these approaches.

6 Further Benchmarking

6.1 Attribute-Based Evaluation

Based on the top-ranked models presented in Tables 7 & 8, we further re-train the top-10[7] deep learning models (using their default settings) on the SOC-Sal_train set (1,800 images) and then test them on the SOC-Sal_test set for attribute-based evaluation. In Table 9, we show the performance on subsets of our dataset characterized by a particular attribute. Due to space limitations, in the following discussion, we only select a few representative attributes for further analysis.

Big object (BO) scenes typically occur when objects are close to the camera, enabling tiny text and patterns to be seen clearly. In this case, models that prefer to focus on local information are seriously misled, leading to a considerable decrease in performance (*e.g.*, 8.6% S_α reduction for BAS, 8.7% reduction for CAGVgg, 11.4% reduction for LDF, and 40.7% reduction for COV) compared with their average performance (Avg.). Among all attributes, BOs are the most difficult for both traditional and deep learning models.

Small objects (SOs) are tricky for some SOD models. Four models (*i.e.*, BAS, CVAE, CAGVgg, and RASNet) encounter performance degradation (*e.g.*, from BAS-0.5% to RASNet-3.6%) because SOs are easily ignored during the downsampling of CNNs. Other models instead have enhanced performance on SOs, but significant reduction in performance on BOs.

Heterogeneous objects (HOs) commonly appear in natural scenes. The performance of all models on HOs improves to some degree, fluctuating from 2.9% to 14.3%. We suspect this is because, as shown in Fig. 6, HO images make up a significant proportion of all datasets, so the models are more familiar with this attribute.

Occlusion (OC) scenes occur when objects are partly obscured. Thus, SOD models must capture global semantics to make up for the incomplete information of objects. As observed, traditional models obtain improved performance compared with their average performance. For deep learning models, in contrast, this situation is reversed.

As can be seen in the last row of Table 9 (average performance of each attribute), *MB* and *SO* have the same S_α score. Moreover, the average scores of *AC* and *SC* are very similar. It seems that existing deep learning based SOD models can effectively address appearance change and shape complexity. Similar to the attributes of *OV* and *OC*, *CL* and *MB* remain challenging for existing methods, generating mid-level (*i.e.*, 0.65< S_α <0.70) S-measure scores.

Table 10
The contribution of our dataset-enhancement strategies.

Method / Metric	$S_\alpha \uparrow$	$E_\xi^{max} \uparrow$	$M \downarrow$
ABP [292]	0.752	0.836	0.097
Our-ABP	**0.769**	**0.842**	**0.093**
EGNet [251]	0.756	0.823	0.105
Our-EGNet	**0.759**	**0.831**	**0.100**
CPDVgg [243]	0.775	0.842	0.090
Our-CPDVgg	**0.789**	**0.850**	**0.087**
CAGVgg [272]	0.748	0.811	0.103
Our-CAGVgg	**0.759**	**0.823**	**0.097**
RASNet [259]	0.832	0.887	0.103
Our-RASNet	**0.841**	**0.897**	**0.096**
CVAE [292]	0.849	0.892	0.089
Our-CVAE	**0.863**	**0.902**	**0.086**
LDF [280]	0.835	0.878	0.103
Our-LDF	**0.845**	**0.891**	**0.097**
R2Net [256]	0.857	0.885	0.084
Our-R2Net	**0.868**	**0.899**	**0.080**
BAS [31]	0.842	0.882	0.092
Our-BAS	**0.856**	**0.895**	**0.086**

Table 11
The contribution of each dataset-enhancement strategy.

Method / Metric	$S_\alpha \uparrow$	$E_\xi^{max} \uparrow$	$M \downarrow$
CVAE [292]	0.849	0.892	0.089
LS	0.851	0.895	0.088
SS	0.852	0.894	0.088
RDA	0.855	0.896	0.086
Our-CVAE	**0.863**	**0.902**	**0.086**

6.2 Comparison with Baselines

We introduce three dataset-enhancement strategies to prevent networks from being overconfident as a result of dataset bias. These include label smoothing, random data augmentation and self-supervised learning. We argue that our strategies can be easily used in existing salient object detection frameworks as general data processing techniques. We thus introduce our strategies to nine benchmark salient object detection models and show the performance in Table 10, where "Our-" represents the benchmark models with our dataset-enhancement strategies. Further, we investigate the contribution of each data-enhancement strategy, and show the performance in Table 11, where we choose CVAE [292] as the base model.

Training & Testing Protocols. We retrain the five models in Table 10 with their corresponding training dataset, *e.g.*, MB [32] for RASNet [259], and DUTS [41] for all the other four models.

[7] DFI mode has only released the test code, so we cannot evaluate it.

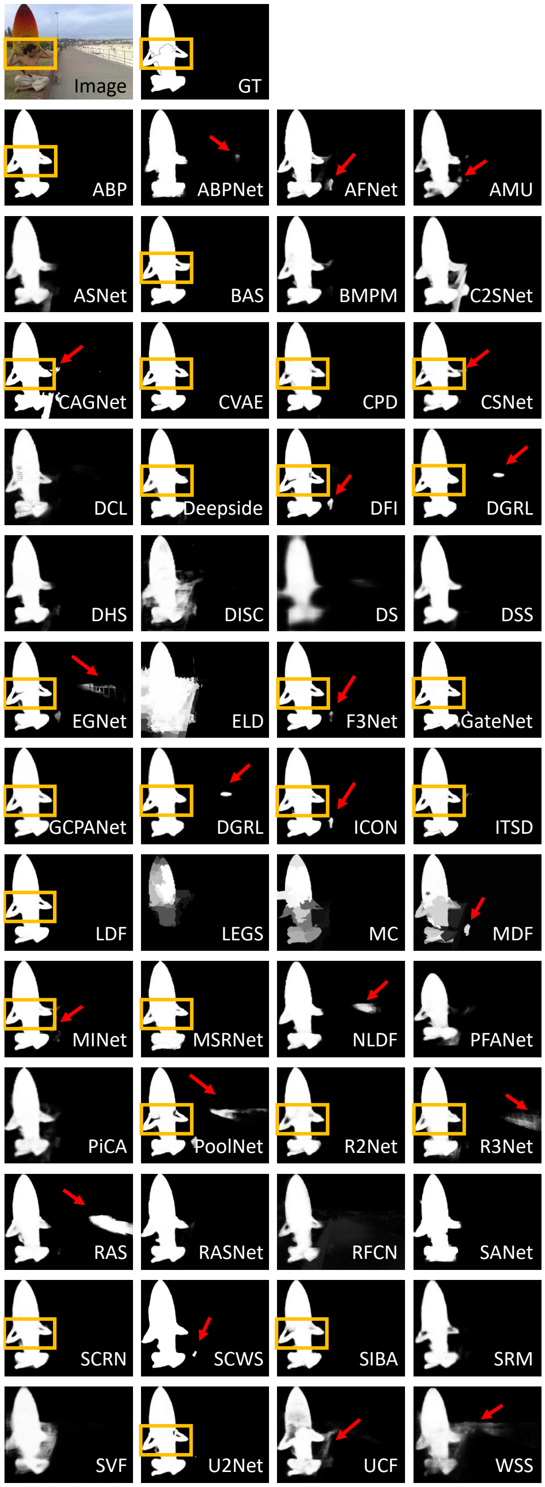


Figure 8. Visualization results of deep learning models.

We follow their original training and testing settings, *e.g.*, same maximum epoch, learning rate, training and testing image sizes.

Discussion. Table 10 shows consistent better performance of models with our strategies, which illustrates effectiveness of our solutions. Further, in Table 11, “LS”, “RDA”, “SS” represent adding label smoothing strategy, random data augmentation and self-supervised learning to the base model respectively. It shows that the random data augmentation achieves the largest performance gain, while label smoothing and self-supervised learning achieves comparable performance improvement. The main reason is that data augmentation introduce diverse samples to the initial training dataset, which is effective in improving model generalization ability. For the self-supervised learning strategy, as the CVAE model [292] has already adopt the multi-scale image as input strategy, we observe slightly improved performance. However, the better performance in general can still validate the effectiveness of the proposed strategy. Label smoothing [294] was introduced to prevent model from over-confidence, thus achieve well-calibrated model. However, there exists no saliency metrics to explain the calibration error of the saliency models. We will investigate in expected calibration error [329] and extend it to saliency detection task in the future to better explain the calibration error issue.

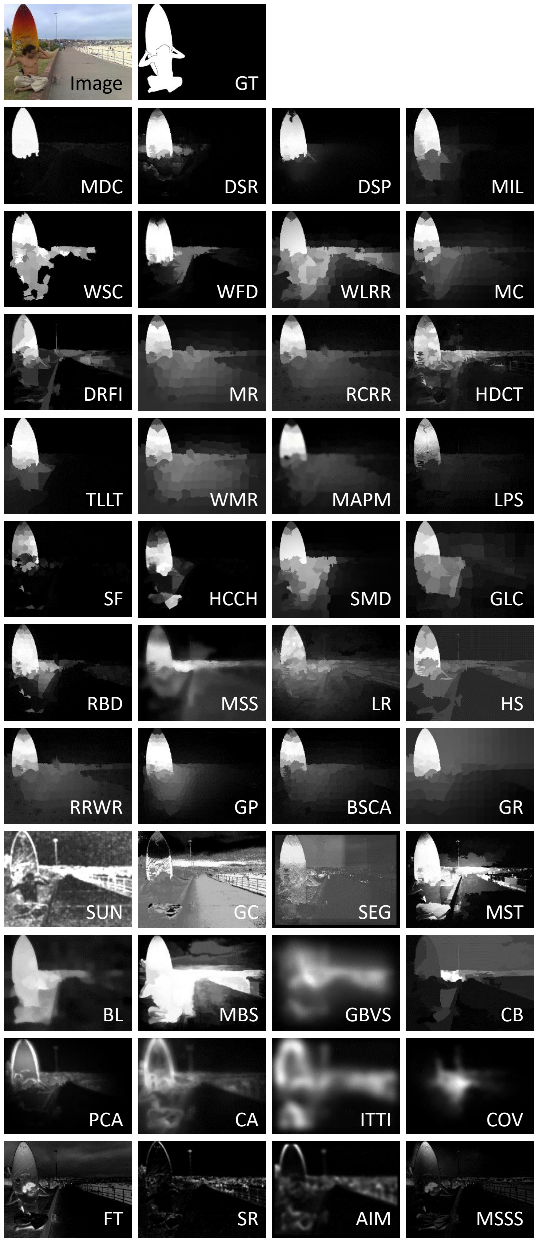


Figure 9. Qualitative results of state-of-the-art traditional approaches.

6.3 Cross-Dataset Generalization

To study the difficulty of existing SOD datasets, we adopt the CDA (cross-data analysis) method [43]. Given N candidate datasets

Table 12
Results for cross-dataset generalization in § 6. UC-Net (CVPR'20) [330] is trained on one dataset and tested on all others. "Sel.": diagonal score (training and testing on the same dataset). "Oth.": mean score on all except for self.

Measure	$S_\alpha \uparrow$ [2]									Drop↓
Train \ Test	SOC	M10K	DU-O	DUTS	ECC	HKU	ILSO	Sel.	Oth.	
SOC [45]	**.884**	.768	.686	.834	.749	.774	.841	.884	.775	12%
M10K [35]	.800	**.921**	.784	.894	.881	.882	.884	.921	.854	7%
DU-O [38]	.833	.898	**.854**	.877	.862	.867	.886	.854	.871	-2%
DUTS [41]	.795	.882	.793	**.910**	.890	.903	.900	.910	.861	5%
ECC [37]	.791	.886	.800	.901	**.901**	.898	.903	.901	.863	4%
HKU [40]	.818	.892	.787	.904	.883	**.910**	.905	.910	.865	5%
ILSO [51]	.841	.888	.790	.898	.882	.896	**.920**	.920	.866	6%
Oth.	.813	.869	.773	.885	.858	.870	.887			

$\{D^n\}_{n=1}^N$, we first train a model on the D_i dataset, and then test it on the other datasets (*i.e.*, $\{D^n\}_{n=1, n\neq i}^N$). Following [46], [331], we randomly select 800 images and 200 images from each dataset as the training set and testing set, respectively.

We train the representative UC-Net [330] (CVPR2020 Best Paper Nomination) on existing popular datasets that contain more than 1,000 images. Table 12 shows the S_α score on each dataset. Each column provides the score of UC-Net tested on a specific dataset and trained on all others. Each row indicates the performance of UC-Net trained on one dataset and tested on all others, demonstrating the generalizability of the dataset adopted for training. We find that when testing on our SOC (*e.g.*, Oth. = 0.813) and DU-O (Oth. = 0.773) datasets, the model performs worse than other datasets. It shows larger differences between SOC/DU-O and the other datasets.

7 Future Directions

Human attention can be influenced by four key factors:

- **Visual properties**. Our attention may be drawn by basic objects' unique visual properties [332].
- **Memory**. If one knows an object well, it is easier for that object to attract one's attention.
- **Goal**. For example, eye fixation records are quite different from attention maps, with a specific goal for viewers.
- **Emotion.** In addition to the above-mentioned factors, we argue that human attention toward the same scene may be affected by one's emotion, *e.g.*, happiness, sadness, anger.

As demonstrated by Cave [332], attentional control is determined by a combination of these factors. Unfortunately, the annotations of existing SOD datasets do not clearly describe which factor they address. Differently, the ground-truth annotations of our SOC are based on the salicon (free-view task) dataset[8], or so-call meaning maps which are used in recent studies [332], [333], [334]. As concluded by Kalash *et al.* [72], the work to date has addressed a relatively ill-posed problem. Thus, we recommend several future directions to re-think SOD tasks at six main research levels:

(1) Data Level: Recently, visual saliency detection tasks have attracted significant interest using 2D (RGB SOD) and 3D (*i.e.*, RGB-D, RGB-T) input data. However, light field SOD (4D), LIDAR SOD, and 360° SOD are still not well-studied. Establishing new datasets for these types of data will largely promote the development of this field. Another interesting avenue for examining saliency detection is to study fine-grained tasks, such as salient instance detection [51], [69], [335], [336] and part-object visual saliency detection [337].

(2) Task Level: Multi-task learning has demonstrated strong performance in recent works [338]. Existing schemes mainly focus on vision tasks, such as joint salient object detection and camouflaged object detection [339], detection of salient objects, edges and skeletons simultaneously [255], and simultaneous detection, ranking, and subitizing of multiple salient objects [70]. With the success of the transformer technique in natural language processing (NLP), introducing multi-modality learning into the saliency detection field may be a feasible way to further incorporate other types of information, such as CV+NLP (similar to [340]), CV+Audio [341], and CV+other modality.

(2) Model Level: A huge number of algorithms have been developed to improve detection accuracy. However, there are several promising directions that could be further studied such as data augmentation techniques [342], efficient SOD models (*e.g.*, lightweight models [284], [343]), new loss functions [287], [344], ranking-based models [70], [138], and transformer-based models [345], [346].

(4) Supervision Level: In addition to the most common fully supervised learning of current SOD models, other supervision strategies, *e.g.*, weakly supervised (*i.e.*, scribble [64], category [347], and polygon), semi-supervised [61], self-supervised [68], [348], and unsupervised [66] learning are also interesting to study.

(5) Evaluation Level: Evaluation metrics are important for model training, testing, and benchmarking. However, the SOD community still utilizes classical metrics such as IoU, F-measure, and MAE. These metrics were designed for universal evaluation rather than for assessing SOD tasks specifically. As a consequence, they do not work well for certain specific applications, such as those with high-quality requirements. We envision that introducing a new metric (*e.g.*, based on the gradient or connectivity error used in [349]) for SOD tasks, such as weighted F-measure [1] and S-measure [2], will be another important research direction.

(6) Application Level: The SOD task belongs to a more general task called class-agnostic object detection (CAOD) [132]. For simple scenes (*e.g.*, those containing only one or two clear objects), SOD is identical to CAOD. From this point of view, SOD models have many potential applications in the real-world (*e.g.*, Alibaba's fashion search system [340]), despite their currently limited number of representative cases [30], [31], [201].

8 Conclusion

In this survey, we identified and addressed the long-ignored *data selection bias* issue in SOD. Different from previous studies, we aimed to explore the SOD task in the wild. To achieve this goal, we collected a new challenging and densely annotated *SOC* dataset; analyzed a large number (∼200) of representative models; conducted the most complete (*i.e.*, 100) benchmarking; devised a series of simple learning strategies to efficiently utilize negative samples and training data; and identified several current challenges and future directions. We hope that these contributions will provide the SOD community an opportunity to explore novel techniques in an open environment. We have tried to cover the most important works. Nevertheless, it is impractical to thoroughly investigate all models in this vast field. We will continue to incorporate new techniques on our website.

References

[1] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *CVPR*, 2014.

[8] http://salicon.net/

[2] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A New Way to Evaluate Foreground Maps," in *ICCV*, 2017.

[3] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment Measure for Binary Foreground Map Evaluation," in *IJCAI*, 2018.

[4] P. Zhang, T. Zhuo, W. Huang, K. Chen, and M. Kankanhalli, "Online object tracking based on cnn with spatial-temporal saliency guided sampling," *Neurocomputing*, vol. 257, pp. 115–127, 2017.

[5] A. Borji, S. Frintrop, D. N. Sihite, and L. Itti, "Adaptive object tracking by learning background context," in *CVPRW*, 2012.

[6] V. Mahadevan and N. Vasconcelos, "On the connections between saliency and tracking," in *NeurIPS*, 2012.

[7] A. Abdulmunem, Y.-K. Lai, and X. Sun, "Saliency guided local and global descriptors for effective action recognition," *Computational Visual Media*, vol. 2, no. 1, pp. 97–106, 2016.

[8] J. He, J. Feng, X. Liu, T. Cheng, T.-H. Lin, H. Chung, and S.-F. Chang, "Mobile product search with bag of hash bits and boundary reranking," in *CVPR*, 2012.

[9] G. Liu and D. Fan, "A model of visual attention for natural image retrieval," in *ISCC-C*, 2013.

[10] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, "Unsupervised object class discovery via saliency-guided multiple class learning," *IEEE TPAMI*, vol. 37, no. 4, pp. 862–875, 2015.

[11] H. Liu, L. Zhang, and H. Huang, "Web-image driven best views of 3d shapes," *TVC*, vol. 28, no. 3, pp. 279–287, 2012.

[12] K. Gu, G. Zhai, X. Yang, W. Zhang, and C. W. Chen, "Automatic contrast enhancement technology with saliency preservation," *IEEE TCSVT*, vol. 25, no. 9, pp. 1480–1494, 2015.

[13] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *CVPR*, 2013.

[14] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," in *ICCV*, 2009.

[15] S. J. Oh, R. Benenson, A. Khoreva, Z. Akata, M. Fritz, and B. Schiele, "Exploiting saliency for object segmentation from image level labels," in *CVPR*, 2017.

[16] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, "Shifting more attention to video salient object detection," in *CVPR*, 2019.

[17] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2photo: Internet image montage," *ACM TOG*, vol. 28, no. 5, p. 124, 2009.

[18] M.-M. Cheng, F.-L. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Repfinder: finding approximately repeated scene elements for image editing," *ACM TOG*, vol. 29, no. 4, p. 83, 2010.

[19] V. Ramanishka, A. Das, J. Zhang, and K. Saenko, "Top-down visual saliency guided by captions," in *CVPR*, 2017.

[20] C. Guo and L. Zhang, "A novel multiresolution spatiotemporal saliency detection model and its applications in image and video compression," *IEEE TIP*, vol. 19, no. 1, pp. 185–198, 2010.

[21] H. Hadizadeh and I. V. Bajic, "Saliency-aware video compression," *IEEE TIP*, vol. 23, no. 1, pp. 19–33, 2014.

[22] Y. Liu, Z. Xu, W. Ye, Z. Zhang, S. Weng, C.-C. Chang, and H. Tang, "Image neural style transfer with preserving the salient regions," *IEEE Access*, vol. 7, pp. 40 027–40 037, 2019.

[23] M.-M. Cheng, X.-C. Liu, J. Wang, S.-P. Lu, Y.-K. Lai, and P. L. Rosin, "Structure-preserving neural style transfer," *IEEE TIP*, vol. 29, pp. 909–920, 2020.

[24] A. Toshev, J. Shi, and K. Daniilidis, "Image matching via saliency region correspondences," in *CVPR*, 2007.

[25] M. J. Islam, R. Wang, K. de Langis, and J. Sattar, "Svam: Saliency-guided visual attention modeling by autonomous underwater robots," *arXiv preprint arXiv:2011.06252*, 2020.

[26] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *CVPR*, 2020.

[27] X. Cheng, H. Xiong, D.-p. Fan, Y. Zhong, M. Harandi, T. Drummond, and Z. Ge, "Implicit motion handling for video camouflaged object detection," in *CVPR*, 2022.

[28] Y. Tu, L. Niu, W. Zhao, D. Cheng, and L. Zhang, "Image cropping with composition and saliency aware aesthetic score map," in *AAAI*, 2020.

[29] J. Kim, T. Misu, Y.-T. Chen, A. Tawari, and J. Canny, "Grounding human-to-vehicle advice for self-driving vehicles," in *CVPR*, 2019.

[30] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C. Lopez, and J. V. Soares, "Leafsnap: A computer vision system for automatic plant species identification," in *ECCV*, 2012.

[31] X. Qin, D.-P. Fan, C. Huang, C. Diagne, Z. Zhang, A. C. Sant'Anna, A. Suàrez, M. Jagersand, and L. Shao, "Boundary-aware segmentation network for mobile and web applications," *arXiv preprint arXiv:2101.04704*, 2021.

[32] T. Liu, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *CVPR*, 2007.

[33] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *CVPR*, 2007.

[34] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *ICCV*, 2001.

[35] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE TPAMI*, vol. 37, no. 3, pp. 569–582, 2015.

[36] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: a benchmark," in *ECCV*, 2012.

[37] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *CVPR*, 2013.

[38] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *CVPR*, 2013.

[39] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *CVPR*, 2014.

[40] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *CVPR*, 2015.

[41] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *CVPR*, 2017.

[42] Z. Yang, S. Soltanian-Zadeh, and S. Farsiu, "Biconnet: An edge-preserved connectivity-based approach for salient object detection," *PR*, vol. 121, p. 108231, 2022.

[43] A. Torralba and A. A. Efros, "Unbiased look at dataset bias," in *CVPR*, 2011.

[44] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *CVPR*, 2019.

[45] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji, "Salient objects in clutter: Bringing salient object detection to the foreground," in *ECCV*, 2018.

[46] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE TPAMI*, 2021.

[47] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *CVPR*, 2009.

[48] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *CVPRW*, 2010.

[49] J. Zhang, S. Ma, M. Sameki, S. Sclaroff, M. Betke, Z. Lin, X. Shen, B. Price, and R. Mech, "Salient object subitizing," in *CVPR*, 2015.

[50] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang, "What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors," in *CVPR*, 2017.

[51] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *CVPR*, 2017.

[52] H. Jiang, M.-M. Cheng, S.-J. Li, A. Borji, and J. Wang, "Joint salient object detection and existence prediction," *Frontiers of Computer Science*, vol. 13, no. 4, pp. 778–788, 2019.

[53] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE TPAMI*, vol. 20, no. 11, pp. 1254–1259, 1998.

[54] A. Borji, "Saliency prediction in the deep learning era: Successes and limitations," *IEEE TPAMI*, vol. 43, no. 2, pp. 679–700, 2021.

[55] T. Liu, J. Sun, N. Zheng, X. Tang, and H. Shum, "Learning to detect a salient object," in *CVPR*, 2007.

[56] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," *IEEE TPAMI*, vol. 33, no. 2, pp. 353–367, 2010.

[57] P. Zhang, W. Liu, Y. Zeng, Y. Lei, and H. Lu, "Looking for the detail and context devils: High-resolution salient object detection," *IEEE TIP*, 2021.

[58] Y. Zeng, P. Zhang, J. Zhang, Z. Lin, and H. Lu, "Towards high-resolution salient object detection," in *ICCV*, 2019.

[59] Q. Zhang, R. Cong, C. Li, M.-M. Cheng, Y. Fang, X. Cao, Y. Zhao, and S. Kwong, "Dense attention fluid network for salient object detection in optical remote sensing images," *IEEE TIP*, 2020.

[60] M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao, "Salient object detection via integrity learning," *arXiv preprint arXiv:2101.07663*, 2021.

[61] Y. Zhou, S. Huo, W. Xiang, C. Hou, and S.-Y. Kung, "Semi-supervised salient object detection using a linear feedback control system model," *IEEE TCYB*, vol. 49, no. 4, pp. 1173–1185, 2019.

[62] G. Li, Y. Xie, and L. Lin, "Weakly supervised salient object detection using image labels," in *AAAI*, 2018.

[63] L. Zhang, J. Zhang, Z. Lin, H. Lu, and Y. He, "Capsal: Leveraging captioning to boost semantics for salient object detection," in *CVPR*, 2019.

[64] J. Zhang, X. Yu, A. Li, P. Song, B. Liu, and Y. Dai, "Weakly-supervised salient object detection via scribble annotations," in *CVPR*, 2020.

[65] Y. Zeng, M. Feng, H. Lu, G. Yang, and A. Borji, "An unsupervised game-theoretic approach to saliency detection," *IEEE TIP*, vol. 27, no. 9, pp. 4545–4554, 2018.

[66] J. Zhang, T. Zhang, Y. Dai, M. Harandi, and R. Hartley, "Deep unsupervised saliency detection: A multiple noisy labeling perspective," in *CVPR*, 2018.

[67] T. Nguyen, M. Dax, C. K. Mummadi, N. Ngo, T. H. P. Nguyen, Z. Lou, and T. Brox, "Deepusps: Deep robust unsupervised saliency prediction via self-supervision," in *NeurIPS*, 2019.

[68] J. Wang, S. Zhu, J. Xu, and D. Cao, "The retrieval of the beautiful: Self-supervised salient object detection for beauty product retrieval," in *ACM MM*, 2019.

[69] R. Fan, M.-M. Cheng, Q. Hou, T.-J. Mu, J. Wang, and S.-M. Hu, "S4net: Single stage salient-instance segmentation," in *CVPR*, 2019.

[70] M. A. Islam, M. Kalash, and N. D. Bruce, "Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects," in *CVPR*, 2018.

[71] S. He, J. Jiao, X. Zhang, G. Han, and R. W. Lau, "Delving into salient object subitizing and detection," in *CVPR*, 2017.

[72] M. Kalash, M. A. Islam, and N. D. Bruce, "Relative saliency and ranking: Models, metrics, data and benchmarks," *IEEE TPAMI*, vol. 43, no. 1, pp. 204–219, 2019.

[73] A. Siris, J. Jiao, G. K. Tam, X. Xie, and R. W. Lau, "Inferring attention shift ranks of objects for image saliency," in *CVPR*, 2020.

[74] A. Borji and L. Itti, "State-of-the-art in visual attention modeling," *IEEE TPAMI*, vol. 35, no. 1, pp. 185–207, 2012.

[75] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE TIP*, vol. 24, no. 12, pp. 5706–5722, 2015.

[76] T. V. Nguyen, Q. Zhao, and S. Yan, "Attentive systems: A survey," *IJCV*, vol. 126, no. 1, pp. 86–110, 2018.

[77] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li, "Salient object detection: A survey," *Computational Visual Media*, vol. 5, no. 2, pp. 117–150, 2019.

[78] D. Zhang, H. Fu, J. Han, A. Borji, and X. Li, "A review of co-saliency detection algorithms: Fundamentals, applications, and challenges," *ACM TIST*, vol. 9, no. 4, pp. 1–31, 2018.

[79] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, "Review of visual saliency detection with comprehensive information," *IEEE TCSVT*, vol. 29, no. 10, pp. 2941–2959, 2018.

[80] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu, "Advanced deep-learning techniques for salient and category-specific object detection: a survey," *IEEE SPM*, vol. 35, no. 1, pp. 84–100, 2018.

[81] A. Borji, "Saliency prediction in the deep learning era: An empirical investigation," *arXiv preprint arXiv:1810.03716*, vol. 10, 2018.

[82] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, "Rethinking rgb-d salient object detection: Models, data sets, and large-scale benchmarks," *IEEE TNNLS*, vol. 32, no. 5, pp. 2075–2089, 2021.

[83] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "Rgb-d salient object detection: A survey," *Computational Visual Media*, pp. 37–69, 2021.

[84] Y. Jiang, T. Zhou, G.-P. Ji, K. Fu, Q. Zhao, and D.-P. Fan, "Light field salient object detection: A review and benchmark," *Computational Visual Media*, 2022.

[85] D.-P. Fan, T. Li, Z. Lin, G.-P. Ji, D. Zhang, M.-M. Cheng, H. Fu, and J. Shen, "Re-thinking co-salient object detection," *IEEE TPAMI*, 2021.

[86] J. Li, J. Su, C. Xia, and Y. Tian, "Distortion-adaptive salient object detection in 360° omnidirectional images," *IEEE Journal of Selected Topics in Signal Processing*, vol. 14, no. 1, pp. 38–48, 2019.

[87] T.-Y. Lin, M. Maire, S. Belongie, J. Hays, P. Perona, D. Ramanan, P. Dollár, and C. L. Zitnick, "Microsoft COCO: Common objects in context," in *ECCV*, 2014.

[88] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *NeurIPS*, 2011.

[89] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE TPAMI*, vol. 23, no. 11, pp. 1222–1239, 2001.

[90] C. Rother, K. Vladimir, and B. Andrew, "Grabcut: interactive foreground extraction using iterated graph cuts," *ACM TOG*, vol. 23, no. 3, pp. 309–314, 2004.

[91] J. Shi and M. Jitendra, "Normalized cuts and image segmentation," *IEEE TPAMI*, vol. 22, no. 8, pp. 888–905, 2000.

[92] Y. Boykov and V. Kolmogorov, "An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision," *IEEE TPAMI*, vol. 26, no. 9, pp. 1124–1137, 2004.

[93] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," in *NeurIPS*, 2007.

[94] X. Hou and L. Zhang, "Saliency detection: A spectral residual approach," in *CVPR*, 2007.

[95] N. Bruce and J. Tsotsos, "Saliency based on information maximization," in *NeurIPS*, 2006.

[96] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *JOV*, vol. 8, no. 7, pp. 32–32, 2008.

[97] Y.-F. Ma and H.-J. Zhang, "Contrast-based image attention analysis by using fuzzy growing," in *ACM MM*, 2003.

[98] R. Achanta, F. Estrada, P. Wils, and S. Süsstrunk, "Salient region detection and segmentation," in *ICCVS*, 2008.

[99] E. Rahtu, J. Kannala, M. Salo, and J. Heikkilä, "Segmenting salient objects from images and videos," in *ECCV*, 2010.

[100] R. Achanta and S. Süsstrunk, "Saliency detection using maximum symmetric surround," in *ICIP*, 2010.

[101] R. Valenti, N. Sebe, T. Gevers *et al.*, "Image saliency by isocentric curvedness and color." in *ICCV*, 2009.

[102] P. L. Rosin, "A simple method for detecting salient regions," *PR*, vol. 42, no. 11, pp. 2363–2371, 2009.

[103] F. Liu and M. Gleicher, "Region enhanced scale-invariant saliency detection," in *ICME*, 2006.

[104] Y. Hu, D. Rajan, and L.-T. Chia, "Robust subspace analysis for detecting visual attention regions in images," in *ACM MM*, 2005.

[105] Z. Yu and H.-S. Wong, "A rule based technique for extraction of visual attention regions based on real-time clustering," *IEEE TMM*, vol. 9, no. 4, pp. 766–784, 2007.

[106] H. Yu, J. Li, Y. Tian, and T. Huang, "Automatic interesting object extraction from images using complementary saliency maps," in *ACM MM*, 2010.

[107] Y. Xie, H. Lu, and M.-H. Yang, "Bayesian saliency via low and mid level cues," *IEEE TIP*, vol. 22, no. 5, pp. 1689–1698, 2012.

[108] E. Erdem and A. Erdem, "Visual saliency estimation by nonlinearly integrating features using region covariances," *JOV*, vol. 13, no. 4, pp. 11–11, 2013.

[109] C. Yang, L. Zhang, and H. Lu, "Graph-regularized saliency detection with convex-hull-based center prior," *IEEE SPL*, vol. 20, no. 7, pp. 637–640, 2013.

[110] N. Tong, H. Lu, L. Zhang, and X. Ruan, "Saliency detection with multi-scale superpixels," *IEEE SPL*, vol. 21, no. 9, pp. 1035–1039, 2014.

[111] H. Peng, B. Li, R. Ji, W. Hu, W. Xiong, and C. Lang, "Salient object detection via low-rank and structured sparse matrix decomposition," in *AAAI*, 2013.

[112] J. Sun, H. Lu, and S. Li, "Saliency detection based on integration of boundary and soft-segmentation," in *ICIP*, 2012.

[113] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *CVPR*, 2011.

[114] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *CVPR*, 2012.

[115] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *CVPR*, 2013.

[116] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *CVPR*, 2014.

[117] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *CVPR*, 2012.

[118] R. Margolin, A. Tal, and L. Zelnik-Manor, "What makes a patch distinct?" in *CVPR*, 2013.

[119] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform," in *CVPR*, 2014.

[120] L. Mai, Y. Niu, and F. Liu, "Saliency aggregation: A data-driven approach," in *CVPR*, 2013.

[121] C. Scharfenberger, A. Wong, K. Fergani, J. S. Zelek, and D. A. Clausi, "Statistical textural distinctiveness for salient region detection in natural images," in *CVPR*, 2013.

[122] R. Liu, J. Cao, Z. Lin, and S. Shan, "Adaptive partial differential equation learning for visual saliency detection," in *CVPR*, 2014.

[123] Z. Jiang and L. S. Davis, "Submodular salient region detection," in *CVPR*, 2013.

[124] K. Shi, K. Wang, J. Lu, and L. Lin, "PISA: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors," in *CVPR*, 2013.

[125] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," in *ICCV*, 2013.

[126] B. Jiang, L. Zhang, H. Lu, C. Yang, and M.-H. Yang, "Saliency detection via absorbing markov chain," in *ICCV*, 2013.

[127] M.-M. Cheng, J. Warrell, W.-Y. Lin, S. Zheng, V. Vineet, and N. Crook, "Efficient salient region detection with soft image abstraction," in *ICCV*, 2013.

[128] K.-Y. Chang, T.-L. Liu, H.-T. Chen, and S.-H. Lai, "Fusing generic objectness and visual saliency for salient object detection," in *ICCV*, 2011.

[129] D. A. Klein and S. Frintrop, "Center-surround divergence of feature statistics for salient object detection," in *ICCV*, 2011.

[130] P. Jiang, H. Ling, J. Yu, and J. Peng, "Salient region detection by ufo: Uniqueness, focusness and objectness," in *ICCV*, 2013.

[131] X. Li, Y. Li, C. Shen, A. Dick, and A. Van Den Hengel, "Contextual hypergraph modeling for salient object detection," in *ICCV*, 2013.

[132] Y. Jia and M. Han, "Category-independent object-level saliency detection," in *ICCV*, 2013.

[133] Y. Lu, W. Zhang, H. Lu, and X. Xue, "Salient object detection using concavity context," in *ICCV*, 2011.

[134] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *ECCV*, 2012.

[135] H. Jiang, J. Wang, Z. Yuan, T. Liu, N. Zheng, and S. Li, "Automatic salient object segmentation based on context and shape prior," in *BMVC*, 2011.

[136] W. Zou, K. Kpalma, Z. Liu, and J. Ronsin, "Segmentation driven low-rank matrix recovery for saliency detection," in *BMVC*, 2013.

[137] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE TPAMI*, vol. 39, no. 4, pp. 818–832, 2016.

[138] L. Zhang, C. Yang, H. Lu, X. Ruan, and M.-H. Yang, "Ranking saliency," *IEEE TPAMI*, vol. 39, no. 9, pp. 1892–1904, 2017.

[139] J. Wang, H. Lu, X. Li, N. Tong, and W. Liu, "Saliency detection via background and foreground seed selection," *Neurocomputing*, vol. 152, pp. 359–368, 2015.

[140] N. Tong, H. Lu, Y. Zhang, and X. Ruan, "Salient object detection via global and local cues," *PR*, vol. 48, no. 10, pp. 3258–3267, 2015.

[141] S. Chen, L. Zheng, X. Hu, and P. Zhou, "Discriminative saliency propagation with sink points," *PR*, vol. 60, pp. 2–12, 2016.

[142] H. Li, H. Lu, Z. Lin, X. Shen, and B. Price, "Inner and inter label propagation: salient object detection in the wild," *IEEE TIP*, vol. 24, no. 10, pp. 3176–3186, 2015.

[143] J. Sun, H. Lu, and X. Liu, "Saliency region detection based on markov absorption probabilities," *IEEE TIP*, vol. 24, no. 5, pp. 1639–1649, 2015.

[144] F. Huang, J. Qi, H. Lu, L. Zhang, and X. Ruan, "Salient object detection via multiple instance learning," *IEEE TIP*, vol. 26, no. 4, pp. 1911–1922, 2017.

[145] Y. Yuan, C. Li, J. Kim, W. Cai, and D. D. Feng, "Reversion correction and regularized random walk ranking for saliency detection," *IEEE TIP*, vol. 27, no. 3, pp. 1311–1322, 2017.

[146] G.-H. Liu and J.-Y. Yang, "Exploiting color volume and color difference for salient region detection," *IEEE TIP*, vol. 28, no. 1, pp. 6–16, 2019.

[147] K. Fu, C. Gong, I. Y.-H. Gu, and J. Yang, "Normalized cut-based saliency detection by adaptive multi-level region merging," *IEEE TIP*, vol. 24, no. 12, pp. 5671–5683, 2015.

[148] X. Huang and Y.-J. Zhang, "300-fps salient object detection via minimum directional contrast," *IEEE TIP*, vol. 26, no. 9, pp. 4243–4254, 2017.

[149] Q. Liu, X. Hong, B. Zou, J. Chen, Z. Chen, and G. Zhao, "Hierarchical contour closure-based holistic salient object detection," *IEEE TIP*, vol. 26, no. 9, pp. 4537–4552, 2017.

[150] Y. Kong, J. Zhang, H. Lu, and X. Liu, "Exemplar-aided salient object detection via joint latent space embedding," *IEEE TIP*, vol. 27, no. 10, pp. 5167–5177, 2018.

[151] S. Huo, Y. Zhou, J. Lei, N. Ling, and C. Hou, "Iterative feedback control-based salient object segmentation," *IEEE TMM*, vol. 20, no. 6, pp. 1350–1364, 2017.

[152] S. Huo, Y. Zhou, W. Xiang, and S.-Y. Kung, "Semisupervised learning based on a novel iterative optimization model for saliency detection," *IEEE TNNLS*, vol. 30, no. 1, pp. 225–241, 2019.

[153] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 fps," in *ICCV*, 2015.

[154] P. Jiang, N. Vasconcelos, and J. Peng, "Generic promotion of diffusion-based salient object detection," in *ICCV*, 2015.

[155] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *CVPR*, 2015.

[156] N. Otsu, "A threshold selection method from gray-level histograms," *IEEE Trans. Syst. Man Cybern. Syst.*, vol. 9, no. 1, pp. 62–66, 1979.

[157] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *CVPR*, 2015.

[158] F. Yang, H. Lu, and Y.-W. Chen, "Human tracking by multiple kernel boosting with locality affinity constraints," in *ACCV*, 2010.

[159] W.-C. Tu, S. He, Q. Yang, and S.-Y. Chien, "Real-time salient object detection with a minimum spanning tree," in *CVPR*, 2016.

[160] C. Li, Y. Yuan, W. Cai, Y. Xia, and D. Dagan Feng, "Robust saliency detection via regularized random walks ranking," in *CVPR*, 2015.

[161] C. Gong, D. Tao, W. Liu, S. J. Maybank, M. Fang, K. Fu, and J. Yang, "Saliency propagation from simple to difficult," in *CVPR*, 2015.

[162] N. Li, B. Sun, and J. Yu, "A weighted sparse coding framework for saliency detection," in *CVPR*, 2015.

[163] Y. Kong, L. Wang, X. Liu, H. Lu, and X. Ruan, "Pattern mining saliency," in *ECCV*, 2016.

[164] Y. Liu, J. Han, Q. Zhang, and L. Wang, "Salient object detection via two-stage graphs," *IEEE TCSVT*, vol. 29, no. 4, pp. 1023–1037, 2019.

[165] Y. Zhou, T. Zhang, S. Huo, C. Hou, and S.-Y. Kung, "Adaptive irregular graph construction-based salient object detection," *IEEE TCSVT*, vol. 30, no. 6, pp. 1569–1582, 2020.

[166] Y. Zhou, A. Mao, S. Huo, J. Lei, and S.-Y. Kung, "Salient object detection via fuzzy theory and object-level enhancement," *IEEE TMM*, vol. 21, no. 1, pp. 74–85, 2019.

[167] X. Lin, Z.-J. Wang, L. Ma, and X. Wu, "Saliency detection via multi-scale global cues," *IEEE TMM*, vol. 21, no. 7, pp. 1646–1659, 2019.

[168] Y. Xu, X. Hong, F. Porikli, X. Liu, J. Chen, and G. Zhao, "Saliency integration: An arbitrator model," *IEEE TMM*, vol. 21, no. 1, pp. 98–113, 2019.

[169] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *CVPR*, 2016.

[170] L. Zhang, J. Sun, T. Wang, Y. Min, and H. Lu, "Visual saliency detection via kernelized subspace ranking with active learning," *IEEE TIP*, vol. 29, pp. 2258–2270, 2020.

[171] X. Huang, Y. Zheng, J. Huang, and Y.-J. Zhang, "50 fps object-level saliency detection via maximally stable region," *IEEE TIP*, vol. 29, pp. 1384–1396, 2020.

[172] R. Strand, K. C. Ciesielski, F. Malmberg, and P. K. Saha, "The minimum barrier distance," *CVIU*, vol. 117, no. 4, pp. 429–437, 2013.

[173] Y.-Y. Zhang, S. Zhang, P. Zhang, H.-Z. Song, and X.-G. Zhang, "Local regression ranking for saliency detection," *IEEE TIP*, vol. 29, pp. 1536–1547, 2020.

[174] M. Everingham, L. Van Gool, C. Williams, J. Winn, and A. Zisserman, "The pascal visual object classes challenge 2012 (voc2012) results (2012)," in *URL http://www. pascal-network. org/challenges/VOC/voc2011/workshop/index. html*, 2011.

[175] S. He, R. W. Lau, W. Liu, Z. Huang, and Q. Yang, "Supercnn: A superpixelwise convolutional neural network for salient object detection," *IJCV*, vol. 115, no. 3, pp. 330–344, 2015.

[176] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *CVPR*, 2015.

[177] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *CVPR*, 2015.

[178] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *CVPR*, 2015.

[179] Y. Yuan, C. Li, J. Kim, W. Cai, and D. D. Feng, "Dense and sparse labeling with multidimensional features for saliency detection," *IEEE TCSVT*, vol. 28, no. 5, pp. 1130–1143, 2016.

[180] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[181] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep image saliency computing via progressive representation learning," *IEEE TNNLS*, vol. 27, no. 6, pp. 1135–1149, 2016.

[182] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," *IEEE TIP*, vol. 25, no. 8, pp. 3919–3930, 2016.

[183] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *ICLR*, 2015.

[184] J. Kim and V. Pavlovic, "A shape-based approach for salient object detection using deep learning," in *ECCV*, 2016.

[185] A. Krizhevsky, S. Ilya, and G. E Hinton, "Imagenet classification with deep convolutional neural networks," in *NeurIPS*, 2012.

[186] Y. Tang and X. Wu, "Saliency detection via combining region-level and pixel-level predictions with cnns," in *ECCV*, 2016.

[187] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *ECCV*, 2016.

[188] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Unconstrained salient object detection via proposal subset optimization," in *CVPR*, 2016.

[189] S. S. Kruthiventi, V. Gudisa, J. H. Dholakiya, and R. Venkatesh Babu, "Saliency unified: A deep architecture for simultaneous eye fixation prediction and salient object segmentation," in *CVPR*, 2016.

[190] M. Jiang, S. Huang, J. Duan, and Q. Zhao, "SALICON: Saliency in context," in *CVPR*, 2015.

[191] J. Kuen, Z. Wang, and G. Wang, "Recurrent attentional networks for saliency detection," in *CVPR*, 2016.

[192] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *ICIP*, 2014.

[193] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: a benchmark and algorithms," in *ECCV*, 2014.

[194] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *CVPR*, 2016.

[195] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *CVPR*, 2016.

[196] P. Hu, B. Shuai, J. Liu, and G. Wang, "Deep level sets for salient object detection," in *CVPR*, 2017.

[197] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *CVPR*, 2017.

[198] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *CVPR*, 2016.

[199] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *CVPR*, 2017.

[200] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *CVPR*, 2009.

[201] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. Torr, "Deeply supervised salient object detection with short connections," in *CVPR*, 2017.

[202] X. Chen, A. Zheng, J. Li, and F. Lu, "Look, perceive and segment: Finding the salient objects in images via two-stream fixation-semantic cnns," in *ICCV*, 2017.

[203] D. Zhang, J. Han, and Y. Zhang, "Supervision by fusion: Towards unsupervised learning of deep salient object detector," in *ICCV*, 2017.

[204] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *ICCV*, 2017.

[205] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *ICCV*, 2017.

[206] S. Chen, B. Wang, X. Tan, and X. Hu, "Embedding attention and residual network for accurate salient object detection," *IEEE TCYB*, 2018.

[207] K. Fu, Q. Zhao, and I. Y.-H. Gu, "Refinet: A deep segmentation assisted refinement network for salient object detection," *IEEE TMM*, vol. 21, no. 2, pp. 457–469, 2018.

[208] C. Cao, Y. Huang, Z. Wang, L. Wang, N. Xu, and T. Tan, "Lateral inhibition-inspired convolutional neural network for visual attention and saliency detection," in *AAAI*, 2018.

[209] X. Hu, L. Zhu, J. Qin, C.-W. Fu, and P.-A. Heng, "Recurrently aggregating deep features for salient object detection," in *AAAI*, 2018.

[210] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R3net: Recurrent residual refinement network for saliency detection," in *AAAI*, 2018.

[211] S. Xie, R. Girshick, P. Dollar, Z. Tu, and K. He, "Aggregated residual transformations for deep neural networks," in *CVPR*, 2017.

[212] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *ECCV*, 2018.

[213] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *ECCV*, 2018.

[214] Y. Zeng, H. Lu, L. Zhang, M. Feng, and A. Borji, "Learning to promote saliency detectors," in *CVPR*, 2018.

[215] M. Amirul Islam, M. Kalash, and N. D. Bruce, "Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects," in *CVPR*, 2018.

[216] W. Wang, J. Shen, X. Dong, and A. Borji, "Salient object detection driven by fixation prediction," in *CVPR*, 2018.

[217] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *CVPR*, 2018.

[218] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, "Detect globally, refine locally: A novel approach to saliency detection," in *CVPR*, 2018.

[219] N. Liu, J. Han, and M.-H. Yang, "Picanet: Learning pixel-wise contextual attention for saliency detection," in *CVPR*, 2018.

[220] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *CVPR*, 2018.

[221] S. Zhou, J. Wang, F. Wang, and D. Huang, "Se2net: Siamese edge-enhancement network for salient object detection," *arXiv preprint arXiv:1904.00048*, 2019.

[222] Z. Li, C. Lang, Y. Chen, J. Liew, and J. Feng, "Deep reasoning with multi-scale context for salient object detection," *arXiv preprint arXiv:1901.08362*, 2019.

[223] S. Jia and N. D. Bruce, "Richer and deeper supervision network for salient object detection," *arXiv preprint arXiv:1901.02425*, 2019.

[224] L. Zhu, J. Chen, X. Hu, C.-W. Fu, X. Xu, J. Qin, and P.-A. Heng, "Aggregating attentional dilated features for salient object detection," *IEEE TCSVT*, vol. 30, no. 10, pp. 3358–3371, 2019.

[225] G. Huang, Z. Liu, L. Van Der Maaten, and K. Q. Weinberger, "Densely connected convolutional networks," in *CVPR*, 2017.

[226] Y. Tang and X. Wu, "Salient object detection using cascaded convolutional neural networks and adversarial learning," *IEEE TMM*, vol. 21, no. 9, pp. 2237–2247, 2019.

[227] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs," *IEEE TPAMI*, vol. 40, no. 4, pp. 834–848, 2018.

[228] Y. Wang, X. Zhao, X. Hu, Y. Li, and K. Huang, "Focal boundary guided salient object detection," *IEEE TIP*, vol. 28, no. 6, pp. 2813–2824, 2019.

[229] T. V. Nguyen, K. Nguyen, and T.-T. Do, "Semantic prior analysis for salient object detection," *IEEE TIP*, vol. 28, no. 6, pp. 3130–3141, 2019.

[230] M. Kampffmeyer, N. Dong, X. Liang, Y. Zhang, and E. P. Xing, "Connnet: A long-range relation-aware pixel-connectivity network for salient segmentation," *IEEE TIP*, vol. 28, no. 5, pp. 2518–2529, 2019.

[231] P. Zhang, W. Liu, H. Lu, and C. Shen, "Salient object detection with lossless feature reflection and weighted structural loss," *IEEE TIP*, vol. 28, no. 6, pp. 3048–3060, 2019.

[232] D. Zhang, J. Han, Y. Zhang, and D. Xu, "Synthesizing supervision for learning deep saliency network without human annotation," *IEEE TPAMI*, vol. 42, no. 7, pp. 1755–1769, 2019.

[233] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, and S. Kwong, "Nested network with two-stream pyramid for salient object detection in optical remote sensing images," *IEEE TGRS*, vol. 57, no. 11, pp. 9156–9166, 2019.

[234] K. Fu, Q. Zhao, I. Y.-H. Gu, and J. Yang, "Deepside: A general deep framework for salient object detection," *Neurocomputing*, vol. 356, pp. 69–82, 2019.

[235] B. Li, Z. Sun, and Y. Guo, "Supervae: Superpixelwise variational autoencoder for salient object detection," in *AAAI*, 2019.

[236] Y. Zhuge, Y. Zeng, and H. Lu, "Deep embedding features for salient object detection," in *AAAI*, 2019.

[237] Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, and Y. Yu, "Multi-source weak supervision for saliency detection," in *CVPR*, 2019.

[238] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, and E. Ding, "A mutual learning method for salient object detection with intertwined multi-supervision," in *CVPR*, 2019.

[239] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *CVPR*, 2019.

[240] M. Feng, H. Lu, and E. Ding, "Attentive feedback network for boundary-aware salient object detection," in *CVPR*, 2019.

[241] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *CVPR*, 2019.

[242] W. Wang, S. Zhao, J. Shen, S. C. Hoi, and A. Borji, "Salient object detection with pyramid attention and salient edges," in *CVPR*, 2019.

[243] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *CVPR*, 2019.

[244] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *CVPR*, 2019.

[245] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-Aware Salient Object Detection," in *CVPR*, 2019.

[246] G. Xavier and B. Yoshua, "Understanding the difficulty of training deep feedforward neural networks," in *AISTATS*, 2010.

[247] Y. Xu, D. Xu, X. Hong, W. Ouyang, R. Ji, M. Xu, and G. Zhao, "Structured modeling of joint deep feature and prediction refinement for salient object detection," in *ICCV*, 2019.

[248] Y. Liu, Q. Zhang, D. Zhang, and J. Han, "Employing deep part-object relationships for salient object detection," in *ICCV*, 2019.

[249] Y. Zeng, Y. Zhuge, H. Lu, and L. Zhang, "Joint learning of saliency detection and weakly supervised semantic segmentation," in *ICCV*, 2019.

[250] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," in *ICCV*, 2019.

[251] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in *ICCV*, 2019.

[252] S. Zhou, J. Wang, J. Zhang, L. Wang, D. Huang, S. Du, and N. Zheng, "Hierarchical u-shape attention network for salient object detection," *IEEE TIP*, vol. 29, pp. 8417–8428, 2020.

[253] Y. Cai, L. Dai, H. Wang, L. Chen, and Y. Li, "A novel saliency detection algorithm based on adversarial learning model," *IEEE TIP*, vol. 29, pp. 4489–4504, 2020.

[254] X. Li, D. Song, and Y. Dong, "Hierarchical feature fusion network for salient object detection," *IEEE TIP*, vol. 29, pp. 9165–9175, 2020.

[255] J.-J. Liu, Q. Hou, and M.-M. Cheng, "Dynamic feature integration for simultaneous detection of salient object, edge, and skeleton," *IEEE TIP*, vol. 29, pp. 8652–8667, 2020.

[256] M. Feng, H. Lu, and Y. Yu, "Residual learning for salient object detection," *IEEE TIP*, vol. 29, pp. 4696–4708, 2020.

[257] L. Zhang, J. Wu, T. Wang, A. Borji, G. Wei, and H. Lu, "A multistage refinement network for salient object detection," *IEEE TIP*, vol. 29, pp. 3534–3545, 2020.

[258] Y. Liu, J. Han, Q. Zhang, and C. Shan, "Deep salient object detection with contextual information guidance," *IEEE TIP*, vol. 29, pp. 360–374, 2020.

[259] S. Chen, X. Tan, B. Wang, H. Lu, X. Hu, and Y. Fu, "Reverse attention-based residual network for salient object detection," *IEEE TIP*, vol. 29, pp. 3763–3776, 2020.

[260] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE TPAMI*, vol. 42, no. 8, pp. 1913–1927, 2020.

[261] H. Li, G. Li, B. Yang, G. Chen, L. Lin, and Y. Yu, "Depthwise nonlocal module for fast salient object detection using a single thread," *IEEE TCYB*, 2020.

[262] J. Li, Z. Pan, Q. Liu, Y. Cui, and Y. Sun, "Complementarity-aware attention network for salient object detection," *IEEE TCYB*, 2020.

[263] H. Li, G. Li, and Y. Yu, "Rosa: Robust salient object detection against adversarial attacks," *IEEE TCYB*, vol. 50, no. 11, pp. 4835–4847, 2019.

[264] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *CVPR*, 2015.

[265] L. Wang, R. Chen, L. Zhu, H. Xie, and X. Li, "Deep sub-region network for salient object detection," *IEEE TCSVT*, 2020.

[266] Z. Tu, Y. Ma, C. Li, J. Tang, and B. Luo, "Edge-guided non-local fully convolutional network for salient object detection," *IEEE TCSVT*, 2020.

[267] X. Hu, C.-W. Fu, L. Zhu, T. Wang, and P.-A. Heng, "Sac-net: Spatial attenuation context for salient object detection," *IEEE TCSVT*, 2020.

[268] Q. Ren, S. Lu, J. Zhang, and R. Hu, "Salient object detection by fusing local and global contexts," *IEEE TMM*, 2020.

[269] Z. Wu, S. Li, C. Chen, A. Hao, and H. Qin, "A deeper look at image salient object detection: Bi-stream network with a small training dataset," *IEEE TMM*, 2020.

[270] J. Li, Z. Pan, Q. Liu, and Z. Wang, "Stacked u-shape network with channel-wise attention for salient object detection," *IEEE TMM*, 2020.

[271] G. Ma, C. Chen, S. Li, C. Peng, A. Hao, and H. Qin, "Salient object detection via multiple instance joint re-learning," *IEEE TMM*, vol. 22, no. 2, pp. 324–336, 2020.

[272] S. Mohammadi, M. Noori, A. Bahri, S. G. Majelan, and M. Havaei, "Cagnet: Content-aware guidance for salient object detection," *PR*, vol. 103, p. 107303, 2020.

[273] B. Zoph, V. Vasudevan, J. Shlens, and Q. V. Le, "Learning transferable architectures for scalable image recognition," in *CVPR*, 2018.

[274] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *PR*, vol. 106, p. 107404, 2020.

[275] C. Wang, S. Dong, X. Zhao, G. Papanastasiou, H. Zhang, and G. Yang, "Saliencygan: Deep learning semisupervised salient object detection in the fog of iot," *IEEE TII*, vol. 16, no. 4, pp. 2667–2676, 2020.

[276] S. Song, H. Yu, Z. Miao, J. Fang, K. Zheng, C. Ma, and S. Wang, "Multi-spectral salient object detection by adversarial domain adaptation," in *AAAI*, 2020.

[277] B. Wang, Q. Chen, M. Zhou, Z. Zhang, X. Jin, and K. Gai, "Progressive feature polishing network for salient object detection," in *AAAI*, 2020.

[278] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," in *AAAI*, 2020.

[279] J. Wei, S. Wang, and Q. Huang, "F^3net: Fusion, feedback and focus for salient object detection," in *AAAI*, 2020.

[280] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *CVPR*, 2020.

[281] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *CVPR*, 2020.

[282] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *CVPR*, 2020.

[283] J. Zhang, J. Xie, and N. Barnes, "Learning noise-aware encoder-decoder from noisy labels by alternating back-propagation for saliency detection," in *ECCV*, 2020.

[284] S.-H. Gao, Y.-Q. Tan, M.-M. Cheng, C. Lu, Y. Chen, and S. Yan, "Highly efficient salient object detection with 100k parameters," in *ECCV*, 2020.

[285] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *ECCV*, 2020.

[286] Y. Liu, M.-M. Cheng, X. Zhang, G.-Y. Nie, and M. Wang, "Dna: Deeply-supervised nonlinear aggregation for salient object detection," *IEEE TCYB*, 2021.

[287] Z. Chen, H. Zhou, J. Lai, L. Yang, and X. Xie, "Contour-aware loss: Boundary-aware learning for salient object segmentation," *IEEE TIP*, vol. 30, pp. 431–443, 2020.

[288] S. Zhou, J. Wang, L. Wang, J. Zhang, F. Wang, D. Huang, and N. Zheng, "Hierarchical and interactive refinement network for edge-preserving salient object detection," *IEEE TIP*, vol. 30, pp. 1–14, 2020.

[289] S. Yu, B. Zhang, J. Xiao, and E. G. Lim, "Structure-consistent weakly supervised salient object detection with local saliency coherence," in *AAAI*, 2021.

[290] M. Ma, C. Xia, and J. Li, "Pyramidal feature shrinking for salient object detection," in *AAAI*, 2021.

[291] B. Xu, H. Liang, R. Liang, and P. Chen, "Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection," in *AAAI*, 2021.

[292] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. Saleh, S. Aliakbarian, and N. Barnes, "Uncertainty inspired rgb-d saliency detection," *IEEE TPAMI*, 2021.

[293] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE TPAMI*, vol. 43, no. 2, pp. 652–662, 2021.

[294] C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, "Rethinking the inception architecture for computer vision," in *CVPR*, 2016.

[295] K. K. Singh and Y. J. Lee, "Hide-and-seek: Forcing a network to be meticulous for weakly-supervised object and action localization," in *ICCV*, 2017.

[296] H. Zhang, M. Cisse, Y. N. Dauphin, and D. Lopez-Paz, "mixup: Beyond empirical risk minimization," in *ICLR*, 2017.

[297] Z. Feng, C. Xu, and D. Tao, "Self-supervised representation learning by rotation feature decoupling," in *CVPR*, 2019.

[298] L. Xie, J. Wang, Z. Wei, M. Wang, and Q. Tian, "Disturblabel: Regularizing cnn on the loss layer," in *CVPR*, 2016.

[299] T. Miyato, S.-i. Maeda, M. Koyama, K. Nakae, and S. Ishii, "Distributional smoothing with virtual adversarial training," in *ICLR*, 2016.

[300] S. Thulasidasan, G. Chennupati, J. Bilmes, T. Bhattacharya, and S. Michalak, "On mixup training: Improved calibration and predictive uncertainty for deep neural networks," in *NeurIPS*, 2019.

[301] S. Wager, W. Fithian, S. Wang, and P. Liang, "Altitude training: Strong bounds for single-layer dropout," in *NeurIPS*, 2014.

[302] J. C. Peterson, R. M. Battleday, T. L. Griffiths, and O. Russakovsky, "Human uncertainty makes classification more robust," in *ICCV*, 2019.

[303] Y. Li, G. Hu, Y. Wang, T. M. Hospedales, N. M. obertson, and Y. Yang, "Differentiable automatic data augmentation," in *ECCV*, 2020.

[304] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *CVPR*, 2019.

[305] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, and S. Yan, "Object region mining with adversarial erasing: A simple classification to semantic segmentation approach," in *CVPR*, 2017.

[306] Z. Zhong, L. Zheng, G. Kang, S. Li, and Y. Yang, "Random erasing data augmentation," in *AAAI*, 2020.

[307] Y.-T. Chang, Q. Wang, W.-C. Hung, R. Piramuthu, Y.-H. Tsai, and M.-H. Yang, "Mixup-cam: Weakly-supervised semantic segmentation via uncertainty regularization," in *BMVC*, 2020.

[308] H. Guo, Y. Mao, and R. Zhang, "Mixup as locally linear out-of-manifold regularization," in *AAAI*, vol. 33, 2019, pp. 3714–3722.

[309] S. Gidaris, P. Singh, and N. Komodakis, "Unsupervised representation learning by predicting image rotations," in *ICLR*, 2018.

[310] X. Zhai, A. Oliver, A. Kolesnikov, and L. Beyer, "S4l: Self-supervised semi-supervised learning," in *ICCV*, 2019.

[311] X. Chen, S. Xie, and K. He, "An empirical study of training self-supervised visual transformers," *arXiv preprint arXiv:2104.02057*, 2021.

[312] X. Wang, K. He, and A. Gupta, "Transitive invariance for self-supervised visual representation learning," in *ICCV*, 2017.

[313] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020.

[314] H. Caesar, J. Uijlings, and V. Ferrari, "Coco-stuff: Thing and stuff classes in context," in *CVPR*, 2018, pp. 1209–1218.

[315] S. Lazebnik, C. Schmid, and J. Ponce, "A sparse texture representation using local affine regions," *IEEE TPAMI*, vol. 27, no. 8, pp. 1265–1278, 2005.

[316] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," in *MIT Technical Report*, 2012.

[317] F. Perazzi, J. Pont-Tuset, B. McWilliams, L. Van Gool, M. Gross, and A. Sorkine-Hornung, "A benchmark dataset and evaluation methodology for video object segmentation," in *CVPR*, 2016.

[318] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," *IEEE TPAMI*, vol. 41, no. 4, pp. 815–828, 2019.

[319] L. Yuan, F. E. Tay, G. Li, T. Wang, and J. Feng, "Revisiting knowledge distillation via label smoothing regularization," in *CVPR*, 2020.

[320] G. Hinton, O. Vinyals, and J. Dean, "Distilling the knowledge in a neural network," in *NeurIPSW*, 2015.

[321] C.-B. Zhang, P.-T. Jiang, Q. Hou, Y. Wei, Q. Han, Z. Li, and M.-M. Cheng, "Delving deep into label smoothing," *IEEE TIP*, vol. 30, pp. 5984–5996, 2021.

[322] C. Godard, O. Mac Aodha, and G. J. Brostow, "Unsupervised monocular depth estimation with left-right consistency," in *CVPR*, 2017.

[323] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, "Image quality assessment: from error visibility to structural similarity," *IEEE TIP*, vol. 13, no. 4, pp. 600–612, 2004.

[324] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for rgb-d salient object detection and beyond," *IEEE TPAMI*, 2021.

[325] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE TPAMI*, vol. 34, no. 10, pp. 1915–1926, 2011.

[326] X. Zhu, C. Tang, P. Wang, H. Xu, M. Wang, J. Chen, and J. Tian, "Saliency detection via affinity graph learning and weighted manifold ranking," *Neurocomputing*, vol. 312, pp. 239–250, 2018.

[327] C. Tang, P. Wang, C. Zhang, and W. Li, "Salient object detection via weighted low rank matrix recovery," *IEEE SPL*, vol. 24, no. 4, pp. 490–494, 2016.

[328] X. Huang and Y. Zhang, "Water flow driven salient object detection at 180 fps," *PR*, vol. 76, pp. 95–107, 2018.

[329] C. Guo, G. Pleiss, Y. Sun, and K. Q. Weinberger, "On calibration of modern neural networks," in *ICML*, 2017.

[330] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes, "Uc-net: uncertainty inspired rgb-d saliency detection via conditional variational autoencoders," in *CVPR*, 2020.

[331] D.-P. Fan, G.-P. Ji, M.-M. Cheng, and L. Shao, "Concealed object detection," *IEEE TPAMI*, 2022.

[332] K. R. Cave, "Finding meaning in eye movements," *Nature Human Behaviour*, vol. 1, no. 10, pp. 709–710, 2017.

[333] J. M. Henderson, T. R. Hayes, C. E. Peacock, and G. Rehrig, "Meaning and attentional guidance in scenes: A review of the meaning map approach," *Vision*, vol. 3, no. 2, p. 19, 2019.

[334] J. M. Henderson and T. R. Hayes, "Meaning-based guidance of attention in scenes as revealed by meaning maps," *Nature Human Behaviour*, vol. 1, no. 10, pp. 743–747, 2017.

[335] X. Tian, K. Xu, X. Yang, B. Yin, and R. W. Lau, "Weakly-supervised salient instance detection," in *BMVC*, 2020.

[336] Y.-H. Wu, Y. Liu, L. Zhang, W. Gao, and M.-M. Cheng, "Regularized densely-connected pyramid network for salient instance segmentation," *IEEE TIP*, vol. 30, pp. 3897–3907, 2021.

[337] Y. Liu, D. Zhang, Q. Zhang, and J. Han, "Part-object relational visual saliency," *IEEE TPAMI*, 2021.

[338] A. R. Zamir, A. Sax, W. Shen, L. J. Guibas, J. Malik, and S. Savarese, "Taskonomy: Disentangling task transfer learning," in *CVPR*, 2018.

[339] A. Li, J. Zhang, Y. Lv, B. Liu, T. Zhang, and Y. Dai, "Uncertainty-aware joint salient object and camouflaged object detection," in *CVPR*, 2021.

[340] M. Zhuge, D. Gao, D.-P. Fan, L. Jin, B. Chen, H. Zhou, M. Qiu, and L. Shao, "Kaleido-bert: Vision-language pre-training on fashion domain," in *CVPR*, 2021.

[341] G. Wang, C. Chen, D.-P. Fan, A. Hao, and H. Qin, "From semantic categories to fixations: A novel weakly-supervised visual-auditory saliency detection approach," in *CVPR*, 2021.

[342] D. V. Ruiz, B. A. Krinski, and E. Todt, "Ida: Improved data augmentation applied to salient object detection," in *SIBGRAPI Conference on Graphics, Patterns and Images*, 2020.

[343] Y. Liu, X.-Y. Zhang, J.-W. Bian, L. Zhang, and M.-M. Cheng, "Samnet: Stereoscopically attentive multi-scale network for lightweight salient object detection," *IEEE TIP*, vol. 30, pp. 3804–3814, 2021.

[344] D.-P. Fan, G.-P. Ji, X. Qin, and M.-M. Cheng, "Cognitive vision inspired object segmentation metric and loss function," *SSI*, 2020.

[345] Y. Mao, J. Zhang, Z. Wan, Y. Dai, A. Li, Y. Lv, X. Tian, D.-P. Fan, and N. Barnes, "Transformer transforms salient object detection and camouflaged object detection," *arXiv preprint arXiv:2104.10127*, 2021.

[346] N. Liu, N. Zhang, K. Wan, L. Shao, and J. Han, "Visual saliency transformer," in *ICCV*, 2021.

[347] H. Zhang, Y. Zeng, H. Lu, L. Zhang, J. Li, and J. Qi, "Learning to detect salient object with multi-source weak supervision," *IEEE TPAMI*, 2021.

[348] X. Zhao, Y. Pang, L. Zhang, H. Lu, and X. Ruan, "Self-supervised representation learning for rgb-d salient object detection," *arXiv preprint arXiv:2101.12482*, 2021.

[349] R. Deora, R. Sharma, and D. S. S. Raj, "Salient image matting," *arXiv preprint arXiv:2103.12337*, 2021.

Deng-Ping Fan received his PhD degree from the Nankai University in 2019. He joined Inception Institute of AI in 2019. He has published about 30 top journal and conference papers such as TPAMI, TIP, CVPR, ICCV, ECCV, *etc*. His research interests include computer vision and visual attention, especially on RGB salient object detection (SOD), RGB-D SOD, Video SOD, Co-SOD. He won the Best Paper Finalist Award at IEEE CVPR 2019, the Best Paper Award Nominee at IEEE CVPR 2020.

Jing Zhang is currently a PhD student with Research School of Electrical, Energy and Materials Engineering, the Australian National University, ACRV, DATA61-CSIRO. She started her Phd degree in 2018. Her main research interests include saliency detection, weakly supervised learning, generative model. She won the Best Student Paper Prize at DICTA 2017, the Best Deep/Machine Learning Paper Prize at APSIPA ASC 2017 and the Best Paper Award Nominee at IEEE CVPR 2020.

Gang Xu is a Ph.D. student in Media Computing Lab at Nankai University. He is supervised by Prof. Ming-Ming Cheng. He received his bachelor's degree from Xidian University in 2018. He has published several top journal and conference papers such as TPAMI, CVPR, *etc*. His research interests include computer vision and machine learning.

Ming-Ming Cheng received his PhD degree from Tsinghua University in 2012. He then he did 2 years research fellow, with Prof. Philip Torr in Oxford. He is a full professor at Nankai University since 2016, leading the Media Computing Lab. His research interests includes computer graphics, machine learning, computer vision, and image processing. He is an Associate Editor of IEEE TIP. He received several research awards, including the ACM China Rising Star Award, the IBM Global SUR Award.

Ling Shao is currently the CEO and the Chief Scientist of the Inception Institute of AI, Abu Dhabi, United Arab Emirates. He is also the Executive Vice President and a Provost of the Mohamed bin Zayed University of Artificial Intelligence. His current research interests include computer vision, machine learning, and medical imaging. Dr. Shao is a fellow of IEEE, IAPR, IET, and BCS. He is an Associate Editor of the IEEE TIP, the IEEE TNNLS, and several other top journals.