Computational Visual Media
https://doi.org/10.1007/s41095-021-0256-2 Vol. 8, No. 4, December 2022, 509–534

Review Article

Light field salient object detection: A review and benchmark

Keren Fu[1], Yao Jiang[1], Ge-Peng Ji[2], Tao Zhou[3] (✉), Qijun Zhao[1], and Deng-Ping Fan[4]

Abstract Salient object detection (SOD) is a long-standing research topic in computer vision with increasing interest in the past decade. Since light fields record comprehensive information of natural scenes that benefit SOD in a number of ways, using light field inputs to improve saliency detection over conventional RGB inputs is an emerging trend. This paper provides the first comprehensive review and a benchmark for light field SOD, which has long been lacking in the saliency community. Firstly, we introduce light fields, including theory and data forms, and then review existing studies on light field SOD, covering ten traditional models, seven deep learning-based models, a comparative study, and a brief review. Existing datasets for light field SOD are also summarized. Secondly, we benchmark nine representative light field SOD models together with several cutting-edge RGB-D SOD models on four widely used light field datasets, providing insightful discussions and analyses, including a comparison between light field SOD and RGB-D SOD models. Due to the inconsistency of current datasets, we further generate complete data and supplement focal stacks, depth maps, and multi-view images for them, making them consistent and uniform. Our supplemental data make a universal benchmark possible. Lastly, light field SOD is a specialised problem, because of its diverse data representations and high dependency on acquisition hardware, so it differs greatly from other saliency detection tasks. We provide nine observations on challenges and future directions, and outline several open issues. All the materials including models, datasets, benchmarking results, and supplemented light field datasets are publicly available at `https://github.com/kerenfu/LFSOD-Survey`.

Keywords light field; salient object detection (SOD); deep learning; benchmarking

1 Introduction

In Google I/O 2021, Google introduced its new technology, Project Starline (`https://blog.google/technology/research/project-starline/`), which combines specialized hardware and computer vision technology to create a "magic window" that can connect two remote persons, making them feel as if they are physically sitting in front of each other during the conversation. Such immersive technology, benefits from *light field* displays, and does not require additional glasses or headsets. The three crucial techniques involved are 3D imaging, real-time data compression, and light field-based 3D displays, which are very challenging but have had breakthroughs according to Google. Salient object detection (SOD) from the light field [1] may also benefit these three stages.

Salient object detection (SOD) [2–4] is a fundamental task in computer vision, aiming to detect and segment conspicuous regions or objects in a scene; light field SOD [5, 6] studies the problem of how to realize SOD using light field data. Numerous applications of SOD cover, e.g., object detection and recognition [7–11], semantic segmentation [12–14], unsupervised video object segmentation [15, 16], multimedia compression [17–20], non-photorealistic rendering [21], re-targeting [22], and human–robot

1 College of Computer Science, Sichuan University, and National Key Laboratory of Fundamental Science on Synthetic Vision, Sichuan University, Chengdu 610065, China. E-mail: K. Fu, fkrsuper@scu.edu.cn; Y. Jiang, yaojiangyj@foxmail.com; Q. Zhao, qjzhao@scu.edu.cn.

2 School of Computer Science, Wuhan University, Wuhan 430072, China. E-mail: gepengai.ji@gmail.com.

3 PCA Lab, School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China. E-mail: taozhou.ai@gmail.com (✉).

4 Computer Vision Lab, ETH Zürich, Zürich, Switzerland. E-mail: dengpfan@gmail.com.

Manuscript received: 2021-07-23; accepted: 2021-10-03

interaction [23, 24]. Generally, the abundant cues and information within the light field help algorithms better identify target objects and improve SOD performance compared to conventional SOD that processes single colour images [25–29].

Light field SOD explores how to detect salient objects using light field data as input. In 3D space, a light field [37] captures all the light rays at every spatial location and in every direction. As a result, it can be viewed as an array of images captured by a grid of cameras. Compared to RGB images captured by a regular camera or depth maps acquired by a depth sensor, the light field data acquired by a plenoptic camera records more comprehensive and complete information about natural scenes, covering, for example, depth information [38–44], focusness cues [5, 42], and angular changes [42, 45]. Therefore, light field data can benefit SOD in a number of ways. Firstly, light fields can be refocused after being acquired [42]. This enables a stack of images focused at different depths to be produced, providing focusness cues that are useful for SOD [46]. Secondly, a light field can provide images of a scene from an array of viewpoints [47]. Such images have abundant spatial parallax and geometric information. Lastly, depth information for a scene is embedded in light field data and can be estimated from a focal stack or multi-view images by different means, as described in Refs. [38–41]. In this sense, RGB-D data can be considered to be a special degenerate case of light field data. Figure 1 shows example results obtained using light field SOD methods on light field data (a focal stack), as well as RGB-D SOD models on depth data.

Although light field data bring great benefits to SOD, and were first considered in 2014 [5], it still remains somewhat under-explored. Specifically, compared to RGB SOD or RGB-D SOD, there are fewer studies on light field SOD. Despite this sparsity of literature, existing models vary in technical frameworks as well as light field datasets used. However, to the best of our knowledge, there is no comprehensive review or benchmark for light field SOD. Although a comparative study was conducted by Zhang et al. [48] in 2015, they only compared the classic light field SOD model proposed by Li et al. [5] to a set of 2D saliency models to demonstrate the effectiveness of incorporating light field knowledge. Besides, the evaluation was conducted on the LFSD dataset, which only contains 100 light field images. Recently, Zhou et al. [49] briefly summarized existing light field SOD models and related datasets. However, their work was mainly focused on RGB-D based SOD, and only a small part was dedicated to reviewing light field SOD, with insufficient consideration of model details and related datasets. Furthermore, they did not benchmark light field SOD models or provide any performance evaluation. Thus, we believe that the lack of a complete review of existing models and datasets may hinder further research in this field.

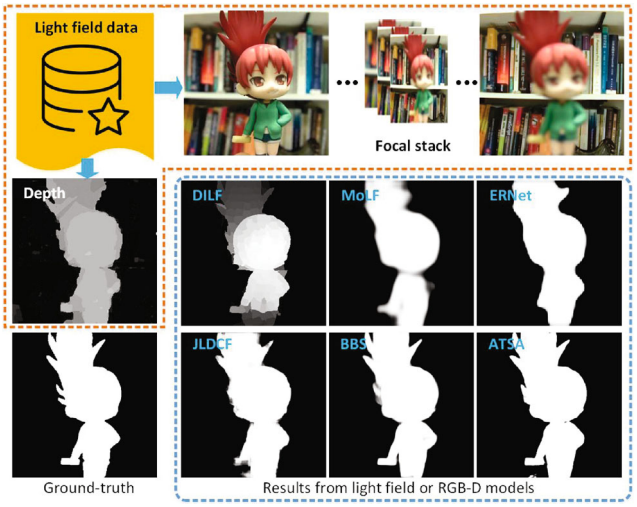


Fig. 1 Salient object detection on a sample scenario using three light field SOD models: DILF [30], MoLF [31], and ERNet [32], and three state-of-the-art RGB-D SOD models: JLDCF [33, 34], BBS [35], and ATSA [36].

Thus, in this paper, we conduct *the first comprehensive review and benchmark for light field SOD*. We review previous studies on light field SOD, including ten traditional models [1, 5, 30, 50–56], seven deep learning-based models [31, 32, 45, 57–60], one comparative study [48], and one brief review [49]. In addition, we also review existing light field SOD datasets [5, 45, 53, 57, 59], and statistically analyze them, covering object size, distance between object and image center, number of focal slices, and number of objects. Due to the inconsistency of datasets (for example, some do not provide focal stacks, while others lack depth maps or multi-view images), we further generate and complete data, including focal stacks, depth maps, and multi-view images for several datasets, to make them consistent and uniform. Furthermore, we benchmark nine light field SOD models [5, 30–32, 45, 50, 54, 58, 59] whose results/code are available,

together with several cutting-edge RGB-D SOD models [33, 35, 36, 61–66], discussing the connection between the two and providing insight into challenges and future directions. All the materials involved in this paper, including collected models, benchmark datasets, results, supplemental light field data, and source code links, are publicly available at `https://github.com/kerenfu/LFSOD-Survey`. The main contributions of this paper are intended to encourage future research in this area, and are four-fold:

- The first systematic review of light field SOD, including models and datasets. Such a survey has long been lacking.
- Analyses of the properties of different datasets. As some lack certain forms of data, e.g., focal stacks, or multi-view images, we generate more data from existing datasets to supplement them, making them complete and uniform.
- A benchmark of nine light field SOD models together with several cutting-edge RGB-D SOD models, using these supplemented datasets, accompanied by insightful discussions.
- An investigation into several challenges for light field SOD and a discussion of its relation to other topics, with directions for future work.

The remainder of the paper is organized as follows. We review light fields, existing models and datasets for light field SOD, with related discussions and analyses in Section 2. In Section 3, we describe evaluation metrics and benchmark results. We then discuss future research directions and outline several open issues in Section 4. Finally, we draw conclusions in Section 5.

2 Preliminaries, models, and datasets

In this section, we first briefly introduce the theory of light fields, its data forms, and how it has been used for SOD. We then review previous works on light field SOD, roughly categorizing them as traditional models and deep learning-based models. Finally, we summarize datasets intended for light field SOD and review their detailed information.

2.1 Light fields

2.1.1 Light fields and light field cameras

A light field [37] consists of all the light rays flowing through every point and in every direction of a 3D space. In 1991, Adelson and Bergen [67] proposed a plenoptic function $P(\theta, \phi, \lambda, t, x, y, z)$ to represent the light field information for wavelength λ at time t in any direction (θ, ϕ) at any point (x, y, z). In an imaging system, the wavelength and time can be represented by RGB channels and different frames, and light usually propagates along a specific path. As a result, Levoy and Hanrahan [68] proposed the two-plane parameterization of the plenoptic function to represent the light field in an imaging system. The two-plane parameterization of the plenoptic function, illustrated in Fig. 2(b), can be formulated as $L(u, v, x, y)$. In this scheme, each ray in the light field is determined by two parallel planes to represent spatial (x, y) and angular (u, v) information. Based on this theory, devices that can capture light fields were invented, commercialised as the Lytro cameras shown in Fig. 2(a). This kind of camera contains the main lens and a micro-lens array placed before the photosensor, where the former serves as the u–v plane, which records the angular information of rays, while the latter serves as the x–y plane, which records the spatial information. Figure 2(b) graphically represents the two-plane parameterization for the light field. Due to the above four-dimensional parameterization, such data are often called 4D light field data [1, 5, 6, 30–32, 45, 48, 50–60].

2.1.2 Forms of light field data

Up to now, all public light field datasets for SOD have been captured by Lytro cameras, the raw data of which are LFP or LFR files (the former are obtained

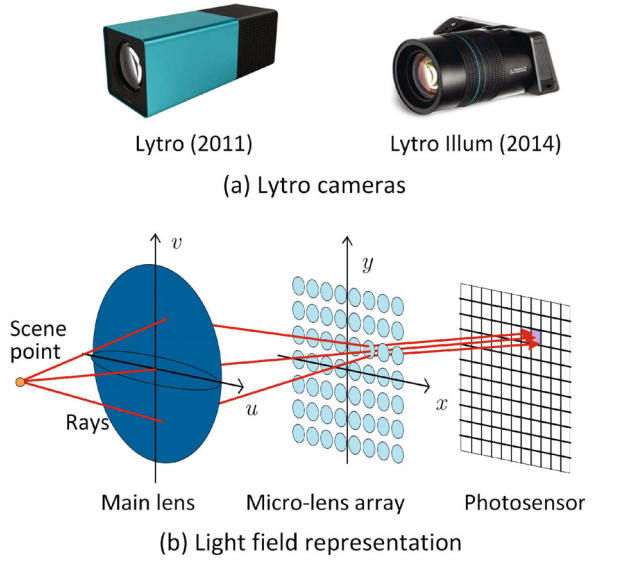


Fig. 2 Lytro cameras (a) and representation of light field (b). For (b), reproduced with permission from Ref. [59], © IEEE 2020.

from Lytro whereas the latter are from Lytro Illum). All images in the current light field datasets were generated by processing LFP or LFR files using Lytro Desktop software (`http://lightfield-forum.com/lytro/lytro-archive/`), or LFToolbox (`http://code.behnam.es/python-lfp-reader/`, or `https://ww2.mathworks.cn/matlabcentral/fileexchange/75250-light-field-toolbox`). Since the raw data cannot be readily utilized, the data forms of light fields used by existing SOD models are diverse, including focal stacks plus all-in-focus images [1, 5, 30–32, 45, 50, 52–55, 58], multi-view images plus center-view images [45, 53, 60], and micro-lens image arrays [51, 59]. As mentioned, depth images can also be synthesized from light field data [38–41], and therefore can form RGB-D data sources for RGB-D SOD models (see Fig. 1). Focal stacks and all-in-focus images are shown in Fig. 3, whereas multi-view images, center-view images, and depth images are shown in Fig. 5.

Specifically, a focal stack (left three columns in Fig. 3) contains a series of images focused at different depths. Such images are generated by processing the raw light field data using digital refocusing techniques [42]. The refocusing principle is demonstrated in Fig. 4, which only shows u and x dimensions. Suppose a light ray enters the main lens at location u, and the imaging plane's position F (the focal distance of the main lens) is changed to F', where $F' = \alpha F$. A refocused image can be computed as follows. First, given the 4D light field L_F, the new light field L_α for the new imaging plane at F' can be derived as

$$L_\alpha(u, v, x, y) = L_F\left(u, v, u + \frac{x-u}{\alpha}, v + \frac{y-v}{\alpha}\right) \tag{1}$$

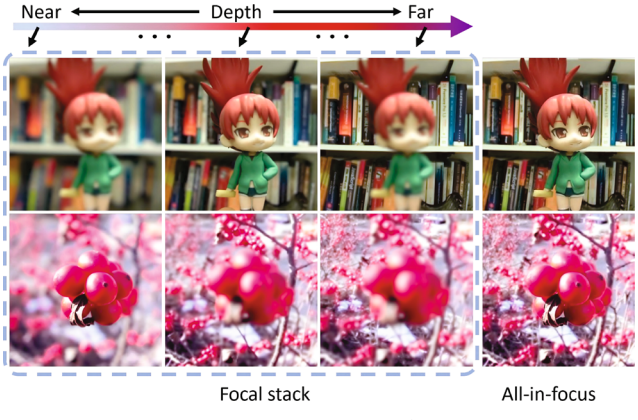


Fig. 3 Focal stacks and all-in-focus images.

Next, after obtaining the new light field $L_\alpha(u, v, x, y)$, a refocused image on the imaging plane can be synthesized as

$$I_\alpha(x, y) = \iint L_\alpha(u, v, x, y)\mathrm{d}u\mathrm{d}v \tag{2}$$

One can see that by changing the parameter α, a series of refocused images can be generated, composing a focal stack. After obtaining the focal stack, an all-in-focus image can be produced by photomontage [69]. For example, an all-in-focus image can be generated by putting all the clear pixels together, where the clarity of pixels can be estimated from associated gradients. It can alternatively be acquired by computing a weighted average of all focus slices. More details can be found in Ref. [70].

In addition to focal stacks, multi-view images (see Fig. 5) can also be derived from light field data. As noted, in the 4D light field representation $L_F(u, v, x, y)$, (u, v) encode angular information about incoming rays. Thus, an image from a certain viewpoint can be generated by sampling in a specific angular direction (u^*, v^*), and the image can be represented by $L_F(u^*, v^*, x, y)$. By varying (u^*, v^*), multi-view images can be synthesized. In particular, when the angular direction (u^*, v^*) is equal to that of the central view, namely (u_0, v_0), the center-view image is achieved. On the other hand, micro-lens images can be generated by sampling the (x, y) dimensions. Providing a micro-lens location (x^*, y^*) leads to a micro-lens image $L_F(u, v, x^*, y^*)$, which captures multiple perspectives of a scene point. Note that by varying (x^*, y^*), different micro-lens images can be obtained, which together compose a micro-lens image array representing complete light field information. Micro-lenses and multi-view images are visualized in Ref. [59].

Moreover, depth maps containing scene depth information can also be estimated from a light field.

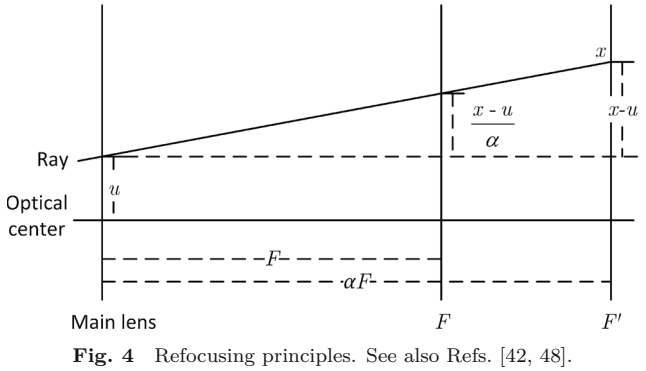


Fig. 4 Refocusing principles. See also Refs. [42, 48].

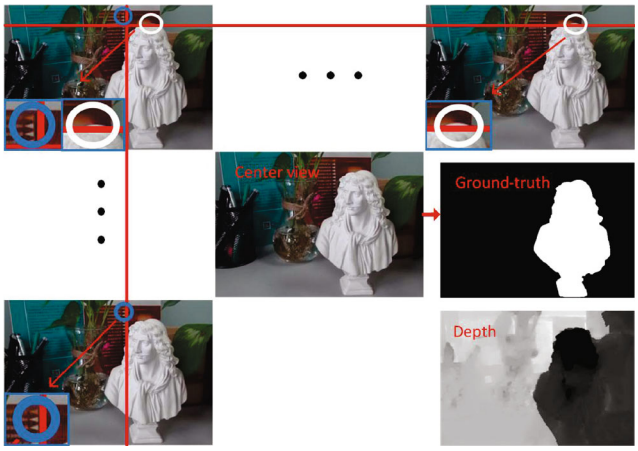


Fig. 5 Multi-view images (including the center-view image), the depth map, and the ground-truth. Note the inconspicuous parallax (disparity) conveyed by the multi-view images (close up at bottom-left in each multi-view image).

Depth information is embedded in the focusness and angular cues, and a depth map can be generated by combining them [38–41, 43, 44].

2.2 Light field SOD models and reviews

We next review and discuss existing models proposed for light field SOD, including ten traditional models that employ hand-crafted features, and seven deep learning-based models. Also, one comparative study and one brief review are revisited. Details of all these works are summarized in Table 1.

2.2.1 Traditional models

As summarized in Table 1, traditional light field SOD models often extend various hand-crafted features and hypotheses which are widely adopted in conventional saliency detection [4], such as global or local color contrast, background priors, and object location cues, to the case of light field data. Some tailored features like focusness, depth, and light-field flow, are also incorporated. Furthermore, these traditional models tend to employ some post-refinement steps, e.g., an optimization framework [1, 30, 51, 53, 55] or CRF [55], to achieve saliency maps with better spatial consistency and more accurate object boundaries. Regarding the data forms used, almost all traditional models work with focal stacks, while depth is incorporated into some of them. Only two traditional models consider using the multi-view [53] and micro-lens data [51]. Further, due to early dataset scarcity, almost all traditional models were evaluated only on the small LFSD dataset constructed by Ref. [5].

Despite early progress made by these traditional models, due to general limitations of hand-crafted features, they hardly generalize well to challenging and complex scenarios compared to modern deep learning models. Here below we briefly review the key features of these traditional models without taxonomy, because they adopt overlapping features but quite diverse computational techniques.

LFS [5] was the earliest work on light field SOD, where the first dataset was also proposed. LFS first incorporated a focusness measure with location priors to determine the background and foreground slices. Then, in the all-in-focus image, it computed the background prior and contrast cues to detect saliency candidates. Finally, a saliency map was generated by incorporating the saliency candidates in the all-in-focus image with those in the foreground slices, where objectness cues were used to weight the candidates. An extension of this work was published in Ref. [6].

WSC [50] was proposed as a unified framework for 2D, 3D, and light field SOD problems; it can handle heterogeneous data. Based on a weighted sparse coding framework, the authors first used a non-saliency dictionary to reconstruct a reference image, where patches with high reconstruction error were selected as the saliency dictionary. This saliency dictionary was later refined by iteratively running the weighted sparse framework to achieve the final saliency map. For light field data, features used for dictionary construction were derived from the all-in-focus RGB image, depth map, and also focal stacks.

DILF [30] computed depth-induced contrast saliency and color contrast saliency from the all-in-focus image and depth image, which were then used to generate a contrast saliency map. It also computed background priors based on a focusness measure embedded in the focal stacks and used them as weights to eliminate background distractions and enhance saliency estimation.

RL [51] proposed to estimate the relative locations of scene points using a filtering process. Such relative locations, which convey scene depth information, were then incorporated with the robust background detection and saliency optimization framework proposed in Ref. [71] to achieve enhanced saliency detection.

BIF [52] used a Bayesian framework to fuse multiple features extracted from RGB images, depth

Table 1 Overview of light field SOD models and review works. FS = focal stacks, DE = depth maps, MV = multi-view images, ML = micro-lens images, OP = open-source. FS, DE, MV, and ML indicate the data form input to a model. New datasets are highlighted in **bold** under Main components

	Model	Pub.	Year	Training dataset(s)	Testing dataset(s)	Main components	FS	DE	MV	ML	OP
Traditional models	LFS [5]	CVPR	2014	—	LFSD	Focusness measure, location priors, contrast cues, background prior, **new dataset (LFSD)**	✓				✓
	WSC [50]	CVPR	2015	—	LFSD	Weighted sparse coding, saliency/non-saliency dictionary construction	✓	✓			✓
	DILF [30]	IJCAI	2015	—	LFSD	Depth-induced/color contrast, background priors by focusness	✓	✓			✓
	RL [51]	ICASSP	2016	—	LFSD	Relative locations, guided filter, micro-lens images				✓	
	BIF [52]	NPL	2017	—	LFSD	Bayesian framework, boundary prior, color/depth-induced contrast	✓	✓			
	LFS [6]	TPAMI	2017	—	LFSD	An extension of Ref. [5]	✓				✓
	MA [53]	TOMM	2017	—	LFSD + HFUT-Lytro	Superpixels intra-cue distinctiveness, light-field flow, **new dataset (HFUT-Lytro)**	✓	✓	✓		
	SDDF [56]	MTAP	2018	—	LFSD	Background priors, gradient operator, color contrast, local binary pattern histograms	✓				
	SGDC [1]	CVPR	2018	—	LFSD	Focusness cues, color, and depth contrast	✓	✓			
	RDFD [54]	MTAP	2020	—	LFSD	Region-based depth feature descriptor, dark channel prior, multi-layer cellular automata	✓				
	DCA [55]	TIP	2020	—	LFSD	Depth-induced cellular automata, object-guided depth	✓	✓			
Deep learning models	DLLF [57]	ICCV	2019	DUTLF-FS	LFSD + DUTLF-FS	VGG-19, attention subnetwork, ConvLSTM, adversarial examples, **new dataset (DUTLF-FS)**	✓				
	DLSD [45]	IJCAI	2019	DUTLF-MV	DUTLF-MV	View synthesis network, multi-view detection/attention, VGG-19, **new dataset (DUTLF-MV)**			✓		✓
	MoLF [31]	NIPS	2019	DUTLF-FS	HFUT-Lytro + LFSD + DUTLF-FS	VGG-19, memory-oriented spatial fusion, memory-oriented feature integration	✓				✓
	ERNet [32]	AAAI	2020	DUTLF-FS + HFUT-Lytro	HFUT-Lytro + LFSD + DUTLF-FS	VGG-19, ResNet-18, multi-focusness recruiting/screening modules, distillation	✓				✓
	LFNet [58]	TIP	2020	DUTLF-FS	HFUT-Lytro + LFSD + DUTLF-FS	VGG-19, refine unit, attention block, ConvLSTM	✓				
	MAC [59]	TIP	2020	Lytro Illum	Lytro Illum + LFSD + HFUT-Lytro	Micro-lens images/image arrays, DeepLab-v2, model angular changes, **new dataset (Lytro Illum)**				✓	✓
	MTCNet [60]	TCSVT	2020	Lytro Illum	Lytro Illum + HFUT-Lytro	Edge detection, depth inference, feature-enhanced salient object generator			✓		
Reviews	CS [48]	NEURO	2015	—	LFSD	Comparative study between 2D vs. light field saliency					
	RGBDS [49]	CVM	2021	—	—	In-depth RGB-D SOD survey, brief review of light field SOD					

maps, and focal stacks. Inspired by image SOD methods, this model utilized a boundary connectivity prior, background likelihood scores, and color contrast to generate background probability maps, foreground slices, color-based saliency maps, and depth-induced contrast maps, which are fused by a two-stage Bayesian scheme.

MA [53] measured the saliency of a superpixel by computing the intra-cue distinctiveness between pairs of superpixels, where features considered included color, depth, and flow inherited from different focal planes and multiple viewpoints. The light-field flow was first employed in this method, estimated from focal stacks and multi-view sequences, to capture depth discontinuities/contrast. The saliency measure was later enhanced using a location prior and a random-search-based weighting strategy. In addition, the authors proposed a new light field SOD dataset, which was the largest at that time.

SDDF [56] made use of depth information embedded in focal stacks to conduct accurate saliency detection. A background measurement was first obtained by applying a gradient operator to focal stack images, and the focal slice with the highest measurement was chosen as the background layer. A coarse prediction was generated by separating the background and foreground in the all-in-focus image using the derived background regions, and the final saliency map was calculated globally from both color and texture (local binary pattern histograms) contrast based on the coarse saliency map.

SGDC [1] presented a contrast-enhanced saliency

detection approach for optimizing a multi-layer light field display. It first computed a superpixel-level focusness map for each refocused image and then chose the refocused image with the highest background likelihood score to derive background cues. These were then incorporated with color and depth contrast saliency. The final results were optimized by the optimization framework in Ref. [71].

RDFD [54] addressed the light field SOD problem via a multiple cue integration framework. A region-based depth feature descriptor (RDFD) defined over the focal stack was proposed, based on the observation that dark channel priors [72] can be used to estimate the degree of defocusing or blur. The RDFD was generated by integrating the degrees of defocusing over all focal stack images, alleviating the limitation of requiring accurate depth maps. RDFD features were used to compute a region-based depth contrast map and a 3D spatial distribution prior. These cues were merged into a single map using a multi-layer cellular automaton.

DCA [55] proposed a depth-induced cellular automata (DCA) for light field SOD. Firstly, it used the focusness and depth cues to calculate an object-guided depth map and select background seeds. Based on the seeds, a contrast saliency map was computed and multiplied by the object-guided depth map to achieve a depth-induced saliency map, which was subsequently optimized by DCA. Finally, the optimized map was combined with the depth-induced saliency map. A Bayesian fusion strategy and CRF were employed to refine the prediction.

2.2.2 Deep learning-based models

Due to the powerful learning ability of deep neural networks, deep learning-based models can achieve superior accuracy and performance [57] over traditional light field SOD models. Another advantage of deep models is that they can directly learn from a large amount of data without hand-crafted feature engineering. Therefore, as shown in Table 1, the scarcity of datasets has been somewhat alleviated in the deep learning era, as three new datasets have been introduced to better train deep neural networks. Still, most deep models take a focal stack as network input. Due to the multi-variable property of focal stacks, modules such as attention mechanisms [31, 32, 45, 57, 58] and ConvLSTMs [31, 32, 57, 58] are preferred. We argue that there may be different ways to classify deep models. A straightforward approach considers what kind of light field data is utilized, as indicated in Table 1. While four models, DLLF [57], MoLF [31], ERNet [32], LFNet [58] resort to focal stacks, DLSD [45] and MTCNet [60] utilize multi-view images, and MAC [59] uses micro-lens images. Different input data forms often lead to different network designs. Note that for DLSD [45], multi-view images processed are indeed rendered from an input single-view image, so this method can be applied to cases no matter whether multi-view images are available or not.

However, since using deep learning-based techniques for light field SOD are the leading trend, in this paper, we divide deep models into five categories according to their architectures, including *late-fusion scheme*, *middle-fusion scheme*, *knowledge distillation-based scheme*, *reconstruction-based scheme*, and *single-stream scheme*: see Fig. 6. Their descriptions and associated models are briefly introduced as follows.

Late-fusion models (Fig. 6(a), DLLF [57], MTCNet [60]) aim to obtain individual predictions from the input focal stack/multi-view images and all-in-focus/center-view image, and then simply fuse the results. Note that late fusion is a classical strategy also widely adopted in previous multi-modal detection (e.g., RGB-D SOD [49], RGB-D semantic segmentation [34, 73]) due to its simplicity and ease of implementation. However, the fusion process is restrained to the last step with relatively simple integrative computation.

DLLF [57] adopted a two-stream fusion framework that explored focal stacks and all-in-focus images separately. In the focal stack stream, DLLF first extracted features from cascaded focal slices through a fully convolutional network. Diverse features from different slices were then integrated by a recurrent attention network, which employed an attention subnetwork and ConvLSTM [74] to adaptively incorporate weighted features of slices and exploit their spatial relevance. The generated map was then combined with another saliency map derived from the all-in-focus image. In addition, to address the limitation of data for training deep networks, a new large dataset was introduced.

MTCNet [60] proposed a two-stream multi-task collaborative network, consisting of a saliency-aware

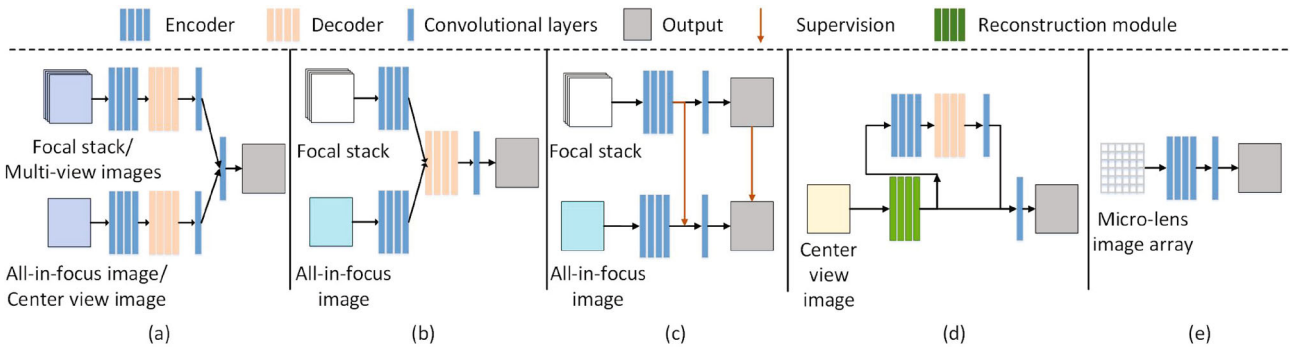


Fig. 6 Architectures of existing deep light field SOD models. (a) Late-fusion: DLLF [57], MTCNet [60]. (b) Middle-fusion: MoLF [31], LFNet [58]. (c) Knowledge distillation-based: ERNet [32]. (d) Reconstruction-based: DLSD [45]. (e) Single-stream: MAC [59]. Here, (a) utilizes the focal stack/multi-view images and all-in-focus/center-view image, while (b, c) utilize the focal stack and all-in-focus image, and (d, e) utilize the center-view image and micro-lens image array.

feature aggregation module (SAFA) and a multi-view inspired depth saliency feature extraction (MVI-DSF) module, to extract representative saliency features with the aid of correlation mechanisms across edge detection, depth inference, and salient object detection. SAFA simultaneously extracted focal-plane, edge, and heuristic saliency features from a center-view image, while MVI-DSF inferred depth saliency features from a set of multi-view images. Finally, MTCNet combined the extracted features using a feature-enhanced operation to obtain the final saliency map.

Middle-fusion models (Fig. 6(b), MoLF [31], LFNet [58]) extract features from the focal stack and all-in-focus image in a two-stream manner. Fusion across intermediate features is then done using an elaborate and complex decoder. Compared to the late-fusion strategy in Fig. 6(a), the main differences are that the features fused are usually hierarchical and intermediate, and the decoder is also a relatively deep convolutional network to mine more complex integration rules.

MoLF [31] featured a memory-oriented decoder that consists of a spatial fusion module (Mo-SFM) and a feature integration module (Mo-FIM), in order to resemble the memory mechanism of human information fusion. Mo-FSM utilized an attention mechanism to learn the importance of different feature maps and a ConvLSTM [74] to gradually refine spatial information. In Mo-FIM, a scene context integration module (SCIM) and ConvLSTM were employed to learn channel attention maps and summarize spatial information.

LFNet [58] proposed a two-stream fusion network to refine complementary information and integrate focusness and blurriness, which change gradually in focal slices. Features extracted from the all-in-focus image and focal stack are fed to a light field refinement module (LFRM) and integration module (LFIM) to generate a final saliency map. In LFRM, features extracted from a single slice were fed to a refinement unit to learn the residuals. In LFIM, an attention block was used to adaptively weight and aggregate slice features.

Knowledge distillation-based methods (Fig. 6(c), ERNet [32]) attempt to transfer focusness knowledge of a teacher network that handles focal stacks, to a student network that processes all-in-focus images. It uses both the features and prediction from the focal stack stream to supervise those features and prediction obtained from the all-in-focus stream, effectively boosting the performance of the latter. In this sense, the student network is actually an RGB SOD network augmented by extra light field knowledge during training.

ERNet [32] consisted of two-stream teacher–student networks based on knowledge distillation. The teacher network used a multi-focusness recruiting module (MFRM) and a multi-focusness screening module (MFSM) to recruit and distil knowledge from focal slices, while the student network took a single RGB image as input for computational efficiency and was enforced to hallucinate multi-focusness features as well as the prediction from the teacher network.

Reconstruction-based schemes (Fig. 6(d), DLSD [45]) focuses on a different aspect as well, namely reconstructing light field data and information from a single input image. This is indeed another

interesting topic as a light field may have various data forms (see Section 2.1.2). With the assistance of the reconstructed light field, an encoder–decoder architecture with a middle- or late-fusion strategy can then be employed to complete the light field SOD. In other words, this scheme is similar to the role of student network in the knowledge distillation-based scheme, being essentially an RGB SOD network augmented by extra light field knowledge during training (in this case, learning to reconstruct light field data).

DLSD [45] treated light field SOD as two sub-problems: light field synthesis from a single-view image and light-field-driven SOD. This model first employed a light field synthesis network, which estimated depth maps in horizontal and vertical directions with two independent convolutional networks. According to the depth maps, the single-view image was warped into horizontal and vertical viewpoints of the light field. A light-field-driven SOD network, consisting of a multi-view saliency detection subnetwork and multi-view attention module, was designed for saliency prediction. Specifically, this model inferred a saliency map from a 2D single-view image, but utilized the light field (the multi-view data) as a middle bridge. To train the model, a new dataset containing multi-view images and pixel-wise ground-truth for the central view was introduced.

Single-stream model (Fig. 6(e), MAC [59]) is inspired by the fact that the light field can be formulated in a single image representation, namely the micro-lens image array [59]. Therefore, unlike Figs. 6(a) and 6(b), this scheme processes the micro-lens image array directly using a single bottom–up stream, without explicit feature fusion.

MAC [59] was an end-to-end deep convolutional network for light field SOD with micro-lens image arrays as input. Firstly, it adopted an MAC (Model Angular Changes) block tailored to model angular changes in individual local micro-lens images and then fed the extracted features to a modified DeepLab-v2 network [75], capturing multiscale information and long-range spatial dependencies. Together with the model, a new Lytro Illum dataset containing high-quality micro-lens image arrays was proposed.

2.2.3 Other reviews

CS [48] provided a comparative study between light field saliency and 2D saliency, showing the advantage of conducting the SOD task on light field data over single 2D images. It compared the classical model LFS [5] with eight 2D saliency models on the LFSD dataset [5]. Five evaluation metrics were used in the paper to show that the light field saliency model achieved better and more robust performance than conventional 2D models.

RGBDS [49] conducted an in-depth and comprehensive survey of RGB-D salient object detection. It reviewed existing RGB-D SOD models from various perspectives, as well as the related benchmark datasets, in detail. As light fields can also provide depth maps, the authors also briefly reviewed light field SOD models and datasets. However, as the main focus of this paper was RGB-D SOD, little space is devoted to reviewing light field SOD, and no associated benchmarking was conducted.

2.3 Light field SOD datasets

2.3.1 Datasets

At present, five datasets exist for the light field SOD task, including LFSD [5], HFUT-Lytro [53], DUTLF-FS [57], DUTLF-MV [45], and Lytro Illum [59]. We summarize details of these datasets in Table 2 and show examples from four datasets (LFSD, HFUT-Lytro, Lytro Illum, and DUTLF-FS) in Fig. 7. A brief introduction to these datasets follows.

LFSD [5] (`https://sites.duke.edu/nianyi/publication/saliency-detection-on-light-field/`) was the first light field dataset collected for SOD, and contains 60 indoor and 40 outdoor scenes. This dataset was captured by a Lytro camera and provides a focal stack, all-in-focus image, depth map, and the corresponding ground-truth for each light field. The image spatial resolution is 360×360. Raw light field data were also available in LFSD. Most images in this dataset contain one single centrally-placed object with a relatively simple background.

HFUT-Lytro [53] (`https://github.com/pencilzhang/MAC-light-field-saliency-net`) contains 255 light fields for both indoor and outdoor scenes. Each light field contains a focal stack with 1–12 slices. The angular resolution is 7×7 and the spatial resolution is 328×328. Focal stacks, all-in-focus images, multi-view images, and coarse depth maps are all provided in this dataset. Several challenges for SOD, e.g., occlusions, cluttered background, and appearance changes, are present in HFUT-Lytro.

Table 2 Overview of light field SOD datasets. MOP = multiple-object proportion (fraction of images in the entire dataset that have more than one object per image), FS = focal stacks, DE = depth maps, MV = multi-view images, ML = micro-lens images, GT = ground-truth, Raw = raw light field data. Here, FS, MV, DE, ML, GT, and Raw indicate the data forms provided by the datasets

Dataset	Number of images	Spatial resolution	Angular resolution	MOP	FS	MV	DE	ML	GT	Raw	Device
LFSD [5]	100 (no official split)	360 × 360	—	0.04	✓		✓		✓	✓	Lytro
HFUT-Lytro [53]	255 (no official split)	328 × 328	7 × 7	0.29	✓	✓	✓		✓		Lytro
DUTLF-FS [57]	1462 (1000 train, 462 test)	600 × 400	—	0.05	✓		✓		✓		Lytro Illum
DUTLF-MV [45]	1580 (1100 train, 480 test)	590 × 400	7 × 7	0.04		✓			✓		Lytro Illum
Lytro Illum [59]	640 (no official split)	540 × 375	9 × 9	0.15				✓	✓	✓	Lytro Illum

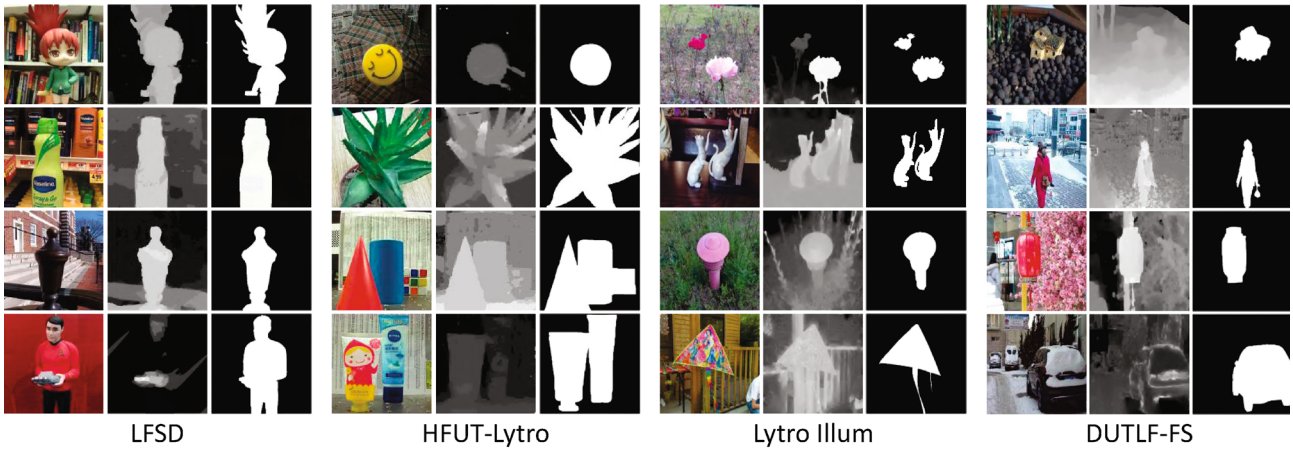


Fig. 7 Examples of RGB images, depth maps, and ground-truth (GT) from four datasets: LFSD [5], HFUT-Lytro [53], Lytro Illum [59], and DUTLF-FS [57]. In each group, RGB images, depth maps, and GT are shown from left to right.

DUTLF-FS [57] (`https://github.com/OIPLab-DUT/ICCV2019\_Deeplightfield\_Saliency`) is one of the largest light field SOD datasets to date, containing 1462 light fields. Both indoor and outdoor scenes were acquired by a Lytro Illum camera. The entire dataset is divided into 1000 training samples and 462 testing samples. All-in-focus images, focal stacks, and the corresponding ground-truth are provided for different light fields. The number of slices in a focal stack ranges from 2 to 12, and the image spatial resolution is 600 × 400. It is worth noting that DUTLF-FS covers various challenges, including different types of objects, low contrast between salient objects and the background, and varied object locations.

DUTLF-MV [45] (`https://github.com/OIPLab-DUT/IJCAI2019-Deep-Light-Field-Driven-Saliency-Detection-from-A-Single-View`) is another large-scale light field dataset for SOD, which was generated from the same database as DUTLF-FS (with 1081 identical scenes). In contrast to other datasets, this dataset was proposed to better exploit angular cues. Therefore, only multi-view images with respect to horizontal and vertical viewpoints are available, together with the ground-truth for the center view image. DUTLF-MV holds 1580 light fields in total, and is divided into training and test sets with 1100 and 480 samples, respectively. The spatial resolution of each image is 400 × 590 and the angular resolution is 7 × 7.

Lytro Illum [59] (`https://github.com/pencilzhang/MAC-light-field-saliency-net`) contains 640 high-quality light fields captured by a Lytro Illum camera. These images vary significantly in object size, texture, background clutter, and illumination. Lytro Illum provides center-view images, micro-lens image arrays, raw light field data, and corresponding ground-truths of the center-view images. The resolution of the micro-lens image arrays is 4860 × 3375, while center-view images and ground-truths are 540 × 375, and the angular resolution can be inferred to be 9 × 9.

2.3.2 Dataset analysis

From the summary in Table 2, we can observe two issues in current datasets, namely limited number of images and non-uniform data representations.

Compared to the large datasets constructed for conventional SOD, such as DUT-OMRON (5168 images) [76], MSRA10K (10,000 images) [2], and DUTS (15,572 images) [77], the existing light field SOD datasets are still small, making it somewhat difficult to evaluate data-driven models and train deep networks. Furthermore, their data forms are not always consistent. For example, Lytro Illum does not provide focal stacks, while DUTLF-FS and DUTLF-MV only provide focal stacks and multi-view images without offering raw data. This makes comprehensive benchmarking very difficult, because a model using focal stacks as input cannot run on DUTLF-MV and Lytro Illum. We will show how we alleviate this problem in Section 3.2, and discuss future directions in Section 4.

To better understand the above-mentioned datasets, we have conducted statistical analyses, including size ratios of salient objects, distributions of normalized object distances from image centers, numbers of focal slices, and numbers of objects. Results are shown in Figs. 8 and 9. Figure 8(a) shows that most objects have size ratios lower than 0.6. HFUT-Lytro and Lytro Illum have relatively small objects, while LFSD has objects that are relatively larger. Figures 8(b) and 9 clearly show the spatial distributions of objects. All five datasets present strong center bias; Fig. 8(b) reveals that objects from Lytro Illum are generally the closest to the image centers.

In addition, numbers of focal slice are summarised in Fig. 8(c). Only three datasets, LFSD, HFUT-Lytro, and DUTLF-FS, provide focal slices. The number of slices varies from 1 to 12 and there are notable differences between different datasets. The distribution peaks for LFSD, HFUT-Lytro, and DUTLF-FS come at 12, 3, and 6 slices respectively. All three datasets have varying numbers of slices, indicating that a light field SOD model using focal stacks should be able to handle differing numbers of input slices. Lastly, from Fig. 8(d), we can see that most images in these datasets have a single object; HFUT-Lytro and Lytro Illum have some images with multiple objects (with higher MOP in Table 2), which could be useful for validating models on detecting multiple objects.

3 Model evaluation and benchmark

In this section, we first review five popular evaluation metrics, and then provide a pipeline for dataset completion. Moreover, we carry out a benchmarking evaluation and provide an analysis of the results.

3.1 Evaluation metrics

In benchmarking light field SOD models, we employ nine widely used metrics, as follows.

Precision–recall curve (PR) [2, 3, 78]. Precision P and recall R are defined as

$$P(T) = \frac{|M^T \cap G|}{|M^T|}, \quad R(T) = \frac{|M^T \cap G|}{|G|} \tag{3}$$

where M^T is a binary mask obtained by thresholding the saliency map with threshold T, and $|\cdot|$ is the total area of the mask. G denotes the ground-truth. A comprehensive precision–recall curve is obtained by changing T from 0 to 255.

F-measure (F_β) [2, 3, 78] is defined as the harmonic-mean of precision and recall:

$$F_\beta = \frac{(1+\beta^2)PR}{\beta^2 P + R} \tag{4}$$

where β is the weight between precision and recall, and β^2 is often set to 0.3 to give more emphasis to precision. Since different F-measure scores can be obtained according to different precision–recall pairs, in this paper, we report the maximum F-measure ($F_\beta^{\max}$) and mean F-measure (F_β^{mean}) computed from

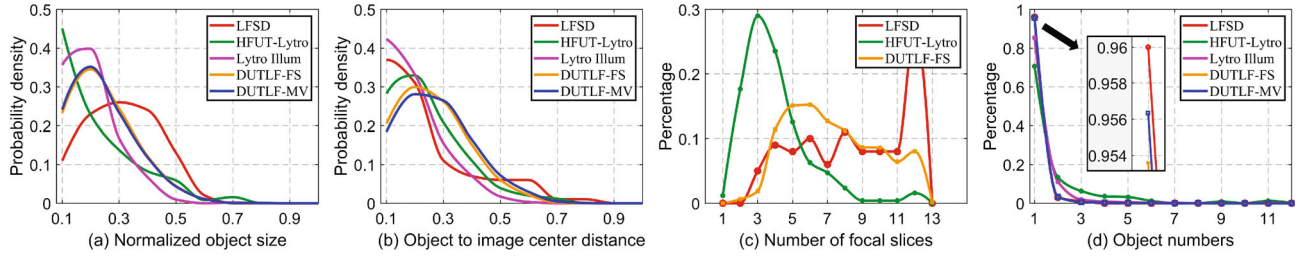


Fig. 8 Statistical summaries of light field datasets, including LFSD [5], HFUT-Lytro [53], Lytro Illum [59], DUTLF-FS [57], and DUTLF-MV [45]. Left to right: distributions of (a) normalized object size, (b) normalized distance between object and image center, (c) number of focal slices, and (d) number of objects.

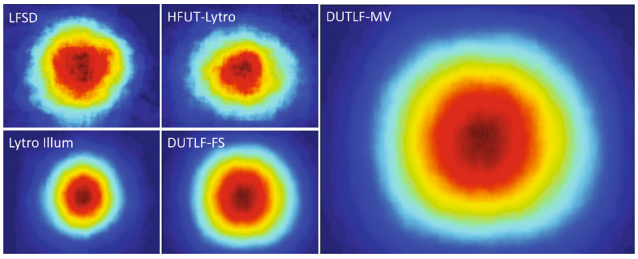


Fig. 9 Object location distribution maps for the five datasets (warmer color means higher probability), computed by averaging ground-truth masks.

the PR curve. We further report the adaptive F-measure (F_β^{adp}) [78], whose threshold is twice the mean of a saliency map.

Mean absolute error (M) [79] is defined as

$$M = \frac{1}{N}\sum_{i=1}^{N}|S_i - G_i| \tag{5}$$

where S_i and G_i denote values at the i-th pixel in the saliency map and ground-truth map. N is the total number of pixels in both maps.

S-measure (S_α) [80, 81] was proposed to measure the spatial structural similarities between the saliency map and ground-truth. It is defined as

$$S_\alpha = \alpha * S_{\text{o}} + (1-\alpha) * S_{\text{r}} \tag{6}$$

where S_{o} and S_{r} denote object-aware and region-aware structure similarity, respectively, and α balances S_{o} and S_{r}. In this paper, we set $\alpha = 0.5$, as recommended in Ref. [80].

E-measure (E_ϕ) [82] is a recently proposed metric which considers both local and global similarity between the prediction and ground-truth. It is defined as

$$E_\phi = \frac{1}{wh}\sum_{i=1}^{w}\sum_{j=1}^{h}\phi(i,j) \tag{7}$$

where $\phi(\cdot)$ denotes the enhanced alignment matrix [82], w and h are the width and height of the ground-truth map respectively, and (i, j) indexes pixels. Since E_ϕ also compares two binary maps, we treat it similarly to the F-measure, thresholding a saliency map with all possible values and reporting the maximum and mean E_ϕ, denoted as $E_\phi^{\max}$ and E_ϕ^{mean} respectively; an adaptive E_ϕ, namely E_ϕ^{adp}, is computed similarly to the adaptive F-measure mentioned above, with threshold twice the mean saliency value [78].

Note that, higher PR curves, F_β, S_α, and E_ϕ, and lower M indicate better performance.

3.2 Dataset completion

As shown in Section 2.3 and Table 2, existing light field SOD datasets face the limitation of having non-uniform data representations. This makes comprehensive benchmarking difficult: due to the lack of specific data, some models cannot be evaluated on certain datasets. To alleviate this issue, we have generated supplementary data for existing datasets, making them complete and uniform, as shown in Table 3, marked by ⭘. This data is on our project site: `https://github.com/kerenfu/LFSOD-Survey` to facilitate future research in this field.

Generally, we can synthesize various data forms using the raw light field data provided by two datasets, i.e., LFSD and Lytro Illum. For Lytro Illum, we generated focal stacks (including all-in-focus images) and depth maps using the Lytro Desktop software. For focal stack generation, we estimated the approximate focus range for each image scene, and then sampled the focal slices within the focus range in equal steps. All-blurred or duplicate slices were removed. The final number of generated focal slices for Lytro Illum ranges from 2 to 16 for each scene, with about 74% of scenes having more than 6 slices. Figure 10 shows an example of the generated focal stack. As mentioned in Section 2.1.2, multi-view images and micro-lens image arrays are generated by angular and spatial sampling of the light field data, respectively. Thus, these two data forms can be transformed into each other. In this way, we generated multi-view images for Lytro Illum from its micro-lens image arrays. We can also synthesize micro-lens image arrays for HFUT-Lytro through the reverse operation. However, we could not synthesize micro-lens image arrays for DUTLF-MV since the authors have only released the multi-view images in the vertical/horizontal direction. By using the raw data, we complemented multi-view images and micro-

Table 3 Dataset completion for light field SOD; compare to Table 2. FS = focal stacks, DE = depth maps, MV = multi-view images, ML = micro-lens images, Raw = raw light field data. ⭘ indicates data that we have completed

Dataset	FS	MV	DE	ML	Raw
LFSD [5]	✓	⭘	✓	⭘	✓
HFUT-Lytro [53]	✓	✓	✓	⭘	
DUTLF-FS [57]	✓		✓		
DUTLF-MV [45]		✓			
Lytro Illum [59]	⭘	⭘	⭘	✓	✓

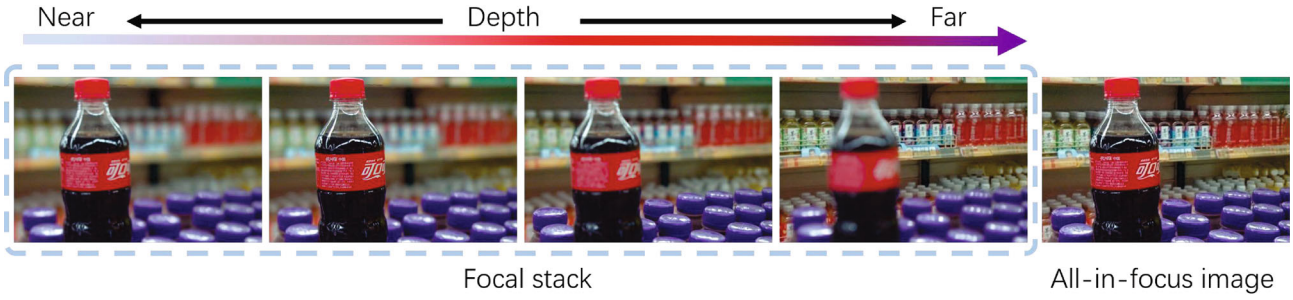


Fig. 10 An example of generated focal slices for Lytro Illum [59], together with the synthesized all-in-focus image.

lens image arrays for LFSD (Fig. 11). The completed data make more comprehensive model evaluation possible. For example, models based on focal stacks, such as MoLF and ERNet, can now be tested on the Lytro Illum dataset. For DUTLF-FS/DUTLF-MV, supplementing the data would be possible in future if the authors were to release the raw (or other) data. If done, DUTLF-FS/DUTLF-MV has the potential to be the *standard training dataset* for future models thanks to its large scale.

3.3 Benchmarking and analysis

3.3.1 Testing

To provide an in-depth understanding of the behaviour of different models, we conducted the first comprehensive benchmarking of nine light field SOD models: LFS [5], WSC [50], DILF [30], RDFD [54] DLSD [45], MoLF [31], ERNet [32], LFNet [58], MAC [59], and nine SOTA RGB-D based SOD models: BBS [35], JLDCF [33, 34], SSF [61], UCNet [62], D3Net [63], S2MA [64], cmMS [65], HDFNet [66], and ATSA [36]) on four existing light field datasets, including the entire LFSD (100 light fields), HFUT-Lytro (255 light fields), Lytro Illum (640 light fields) datasets, and the test set (462 light fields) of DUTLF-FS. Sample images from these datasets are shown in Fig. 7.

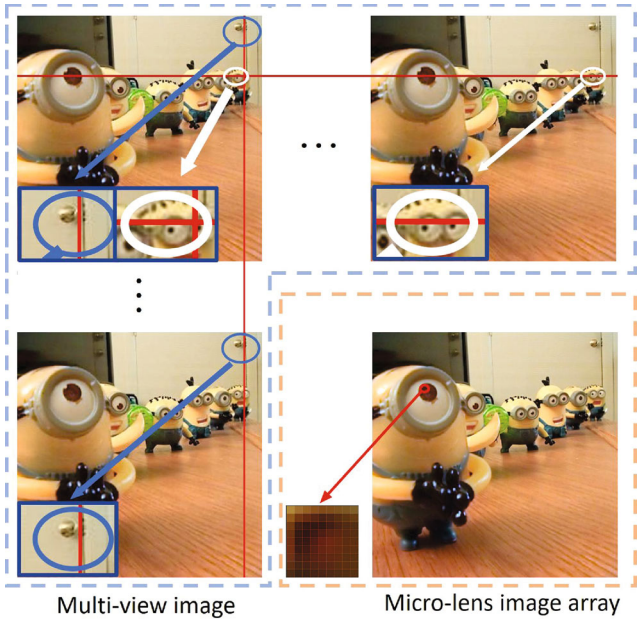


Fig. 11 An example of generated multi-view images (360 × 360) and the micro-lens image array (1080 × 1080) from the LFSD dataset [5]. The bottom-left of each image shows close up details to better reflect the parallax. The micro-lens image array is composed of many micro-lens images [59].

Note the RGB-D SOD models benchmarked here were those that came top in a recent survey [49] and also the latest open-source models in ECCV-2020. All depth maps fed to each model were optionally reversed on an entire dataset to fit the best performance of this model. All benchmarked models have either publicly available source or executable code, or results provided by the authors (authors of RDFD [54] and LFNet [58] have sent us their saliency map results). The nine evaluation metrics described previously: PR, S-measure, max/mean F-measure, max/mean E-measure, adaptive F-measure and E-measure, mean absolute error were used, and the results are reported in Table 4. PR curves, max F-measure curves, and visual comparisons are shown in Figs. 12–15.

Evaluation was not conducted on the DUTLF-MV dataset [45] since it only provides multi-view images, which are incompatible with the input data forms of most light field SOD models. Furthermore, DLSD [45] was not tested on the DUTLF-FS test set because it used about 36% of the testing images from this dataset for training. Also, MAC [59] was not evaluated on Lytro Illum since the authors conducted five-fold cross-validation on this dataset, and so it is not directly comparable to other models. As DUTLF-FS has no micro-lens image arrays after dataset completion (see Table 3), and the quality of micro-lens

Table 4 Quantitative measures: S-measure (S_α) [80], max F-measure ($F_\beta^{\max}$), mean F-measure (F_β^{mean}) [78], adaptive F-measure (F_β^{adp}) [78], max E-measure ($E_\phi^{\max}$), mean E-measure (E_ϕ^{mean}) [82], adaptive E-measure (E_ϕ^{adp}) [78], and MAE (M) [79] of nine light field SOD models (i.e., LFS [5], WSC [50], DILF [30], RDFD [54], DLSD [45], MoLF [31], ERNet [32], LFNet [58], MAC [59]) and nine SOTA RGB-D based SOD models (i.e., BBS [35], JLDCF [33], SSF [61], UCNet [62], D3Net [63], S2MA [64], cmMS [65], HDFNet [66], and ATSA [36]). Light field SOD models are marked by †. N/T indicates that a model was not tested. The best three models for light field and RGB-D based SOD models are highlighted in red, blue, and green, separately. ↑/↓ denotes that a larger/smaller value is better

		Traditional				Deep learning-based													
	Metric	LFS† [5]	WSC† [50]	DILF† [30]	RDFD† [54]	DLSD† [45]	MoLF† [31]	ERNet† [32]	LFNet† [58]	MAC† [59]	BBS [35]	JLDCF [33]	SSF [61]	UCNet [62]	D3Net [63]	S2MA [64]	cmMS [65]	HDFNet [66]	ATSA [36]
LFSD [5]	S_α ↑	0.681	0.702	0.811	0.786	0.786	0.825	0.831	0.820	0.789	0.864	0.862	0.859	0.858	0.825	0.837	0.850	0.846	0.858
	$F_\beta^{\max}$ ↑	0.744	0.743	0.811	0.802	0.784	0.824	0.842	0.824	0.788	0.858	0.867	0.868	0.859	0.812	0.835	0.858	0.837	0.866
	F_β^{mean} ↑	0.513	0.722	0.719	0.735	0.758	0.800	0.829	0.794	0.753	0.842	0.848	0.862	0.848	0.797	0.806	0.850	0.818	0.856
	F_β^{adp} ↑	0.735	0.743	0.795	0.802	0.779	0.810	0.831	0.806	0.793	0.840	0.827	0.862	0.838	0.788	0.803	0.857	0.818	0.852
	$E_\phi^{\max}$ ↑	0.809	0.789	0.861	0.851	0.859	0.880	0.884	0.885	0.836	0.900	0.902	0.901	0.898	0.863	0.873	0.896	0.880	0.902
	E_ϕ^{mean} ↑	0.567	0.753	0.764	0.758	0.819	0.864	0.879	0.867	0.790	0.883	0.894	0.890	0.893	0.850	0.855	0.881	0.869	0.899
	E_ϕ^{adp} ↑	0.773	0.788	0.846	0.834	0.852	0.879	0.882	0.882	0.839	0.889	0.882	0.896	0.890	0.853	0.863	0.890	0.872	0.897
	M ↓	0.205	0.150	0.136	0.136	0.117	0.092	0.083	0.092	0.118	0.072	0.070	0.067	0.072	0.095	0.094	0.073	0.086	0.068
HFUT-Lytro [53]	S_α ↑	0.565	0.613	0.672	0.619	0.711	0.742	0.778	0.736	0.731	0.751	0.789	0.725	0.748	0.749	0.729	0.723	0.763	0.772
	$F_\beta^{\max}$ ↑	0.427	0.508	0.601	0.533	0.624	0.662	0.722	0.657	0.667	0.676	0.727	0.647	0.677	0.671	0.650	0.626	0.690	0.729
	F_β^{mean} ↑	0.323	0.493	0.513	0.469	0.594	0.639	0.709	0.628	0.620	0.654	0.707	0.639	0.672	0.651	0.623	0.617	0.669	0.706
	F_β^{adp} ↑	0.427	0.485	0.530	0.518	0.592	0.627	0.706	0.615	0.638	0.654	0.677	0.636	0.675	0.647	0.588	0.636	0.653	0.689
	$E_\phi^{\max}$ ↑	0.637	0.695	0.748	0.712	0.784	0.812	0.841	0.799	0.797	0.801	0.844	0.778	0.804	0.797	0.777	0.784	0.801	0.833
	E_ϕ^{mean} ↑	0.524	0.684	0.657	0.623	0.749	0.790	0.832	0.777	0.733	0.765	0.825	0.763	0.793	0.773	0.756	0.746	0.788	0.819
	E_ϕ^{adp} ↑	0.666	0.680	0.693	0.691	0.755	0.785	0.831	0.770	0.772	0.804	0.811	0.781	0.810	0.789	0.744	0.779	0.789	0.810
	M ↓	0.221	0.154	0.150	0.214	0.111	0.094	0.082	0.092	0.107	0.089	0.075	0.100	0.090	0.091	0.112	0.097	0.095	0.084
Lytro Illum [59]	S_α ↑	0.619	0.709	0.756	0.738	0.788	0.834	0.843	N/T	N/T	0.879	0.890	0.872	0.865	0.869	0.853	0.881	0.873	0.883
	$F_\beta^{\max}$ ↑	0.545	0.662	0.697	0.696	0.746	0.820	0.827	N/T	N/T	0.850	0.878	0.850	0.843	0.843	0.823	0.857	0.855	0.875
	F_β^{mean} ↑	0.385	0.646	0.604	0.624	0.713	0.766	0.800	N/T	N/T	0.829	0.848	0.836	0.827	0.818	0.788	0.839	0.823	0.848
	F_β^{adapt} ↑	0.547	0.639	0.659	0.679	0.720	0.747	0.796	N/T	N/T	0.828	0.830	0.835	0.824	0.813	0.778	0.835	0.823	0.842
	$E_\phi^{\max}$ ↑	0.721	0.804	0.830	0.816	0.871	0.908	0.911	N/T	N/T	0.913	0.931	0.913	0.910	0.909	0.895	0.914	0.913	0.929
	E_ϕ^{mean} ↑	0.546	0.791	0.726	0.738	0.830	0.882	0.900	N/T	N/T	0.900	0.919	0.907	0.904	0.894	0.873	0.907	0.898	0.919
	E_ϕ^{adapt} ↑	0.771	0.797	0.812	0.815	0.855	0.876	0.900	N/T	N/T	0.912	0.914	0.917	0.907	0.907	0.878	0.915	0.904	0.917
	M ↓	0.197	0.115	0.132	0.142	0.086	0.065	0.056	N/T	N/T	0.047	0.042	0.044	0.048	0.050	0.063	0.045	0.051	0.041
DUTLF-FS [57]	S_α ↑	0.585	0.656	0.725	0.658	N/T	0.887	0.899	0.878	0.804	0.894	0.905	0.908	0.870	0.852	0.845	0.906	0.868	0.905
	$F_\beta^{\max}$ ↑	0.533	0.617	0.671	0.599	N/T	0.903	0.908	0.891	0.792	0.884	0.908	0.915	0.864	0.840	0.829	0.906	0.857	0.915
	F_β^{mean} ↑	0.358	0.607	0.582	0.538	N/T	0.855	0.891	0.843	0.746	0.867	0.885	0.907	0.854	0.820	0.806	0.893	0.841	0.899
	F_β^{adapt} ↑	0.525	0.617	0.663	0.599	N/T	0.843	0.885	0.831	0.790	0.872	0.874	0.903	0.850	0.826	0.791	0.887	0.835	0.893
	$E_\phi^{\max}$ ↑	0.711	0.788	0.802	0.774	N/T	0.939	0.949	0.930	0.863	0.923	0.943	0.946	0.909	0.891	0.883	0.936	0.898	0.943
	E_ϕ^{mean} ↑	0.511	0.759	0.695	0.686	N/T	0.921	0.943	0.912	0.806	0.908	0.932	0.939	0.904	0.874	0.866	0.928	0.889	0.938
	E_ϕ^{adapt} ↑	0.742	0.787	0.813	0.782	N/T	0.923	0.943	0.913	0.872	0.924	0.930	0.942	0.905	0.895	0.870	0.931	0.895	0.936
	M ↓	0.227	0.151	0.156	0.191	N/T	0.051	0.039	0.054	0.102	0.054	0.043	0.036	0.059	0.069	0.079	0.041	0.065	0.039

image arrays in HFUT-Lytro is fairly low due to the low-quality multi-view images, we followed Ref. [59] and instead tested MAC on single up-sampled all-in-focus images from these two datasets. In addition, for ERNet [32], we only evaluated the teacher model since its pre-trained student model is not publicly available. A comprehensive analysis of the results now follows.

3.3.2 Traditional vs. deep models

Compared to the four traditional models shown in Table 1, the deep learning-based SOD models clearly have provide better results on all datasets. The best traditional model evaluated, namely DILF, is generally inferior to any deep light field model. This confirms the power of deep neural networks when applied to this task.

3.3.3 Deep learning models

As shown in Table 1, MoLF, ERNet, and LFNet adopt focal stacks and all-in-focus images as input data, while DLSD and MAC use center-view images and micro-lens image arrays. From Table 4 and Fig. 12, it is clear that MoLF, ERNet, and LFNet are better than DLSD and MAC. It is also worth noting that MoLF and ERNet are the best two methods, probably because they were trained on the large-scale DUTLF-FS dataset with 1000 light fields, with superior network structures. These results also indicate that models based on multi-view or micro-lens images are not as effective as those based on focal stacks. This is probably because that the former are less studied, and the effectiveness of multi-view and micro-lens images is still underexplored. Moreover,

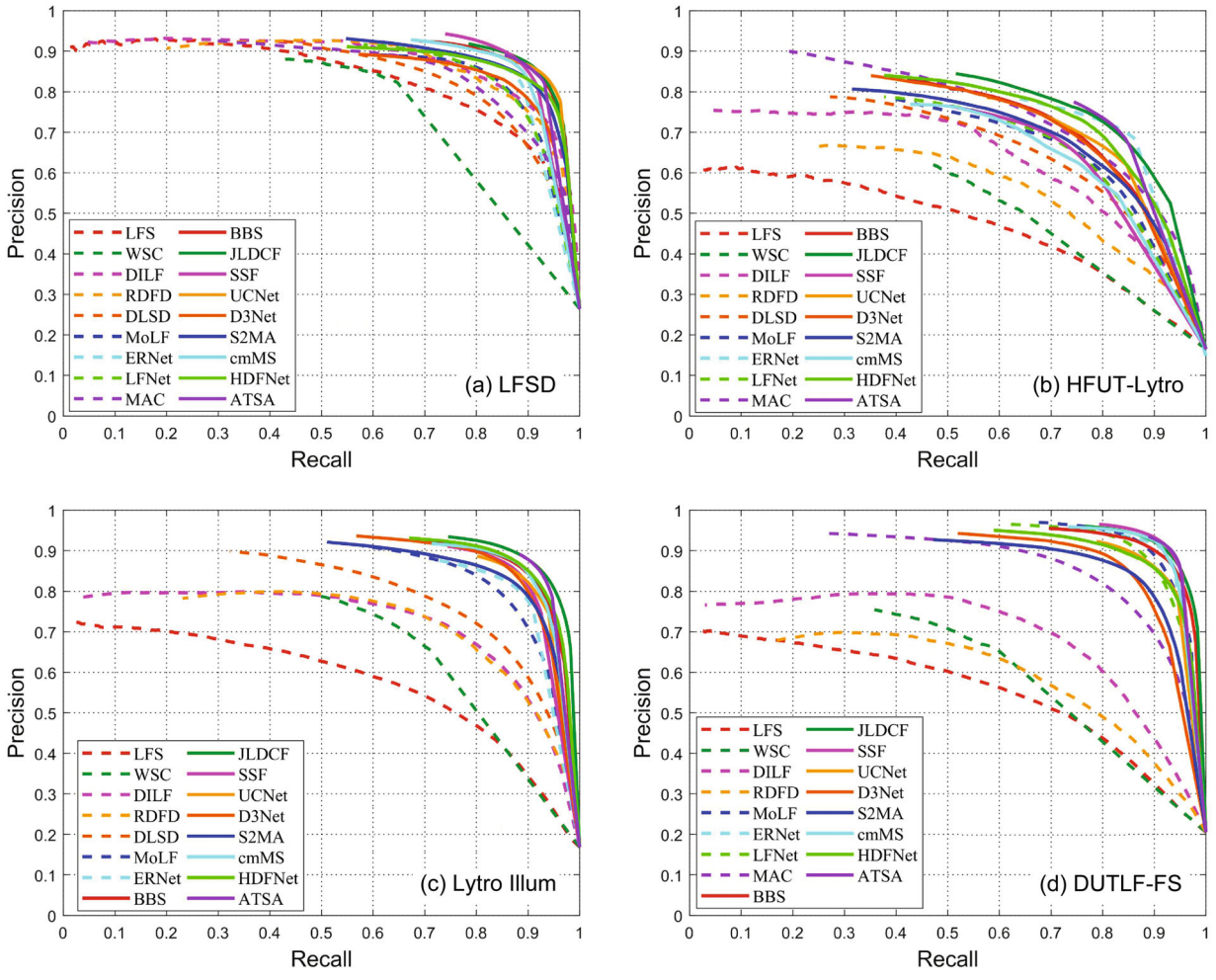


Fig. 12 PR curves for four datasets: (a) LFSD [5], (b) HFUT-Lytro [53], (c) Lytro Illum [59], and (d) DUTLF-FS [57], for nine light field SOD models: LFS [5], WSC [50], DILF [30], RDFD [54], DLSD [45], MoLF [31], ERNet [32], LFNet [58], MAC [59], and nine SOTA RGB-D based SOD models: BBS [35], JLDCF [33, 34], SSF [61], UCNet [62], D3Net [63], S2MA [64], cmMS [65], HDFNet [66], and ATSA [36]. *Solid lines* and *dashed lines* represent the PR curves of *RGB-D based SOD models* and *light field SOD models*, respectively.

the training data may also matter because MAC was trained only on Lytro Illum, which is about half the scale of DUTLF-FS. Among the above five models compared, ERNet gave best accuracy.

3.3.4 Light field and RGB-D SOD models

From the quantitative results in Table 4 and Fig. 12, it can be observed that, the latest cutting-edge RGB-D models achieve comparable or even better results than the light field SOD models. In particular, JLDCF, SSF, and ATSA, are generally better than ERNet on most datasets. The underlying reasons may be two-fold. Firstly, RGB-D based SOD has recently drawn extensive research interest and many powerful and elaborate models have been proposed. Inspired by previous research on the RGB SOD problem [28, 83, 84], these models often pursue edge-preserving results from deep neural networks and employ functional modules and architectures, such as a boundary supplement unit [61], a multi-scale feature aggregation module [36], or a UNet-shaped bottom–up/top–down architecture [33, 64, 85]. In contrast, light field SOD has been less explored and the models and architectures evolve slowly. Edge-aware properties have not yet been considered by most existing models. For example, although the attention mechanism and ConvLSTM are adopted in ERNet, no UNet-like top–down refinement is used to generate edge-aware saliency maps. As evidenced in Figs. 1 and 14, the RGB-D SOD models tend to detect more accurate boundaries than existing deep light field SOD models. Secondly, another reason could be that RGB-D SOD models are trained on more data. For instance, the universally agreed training set for the RGB-D SOD task contains 2200 RGB-D scenes

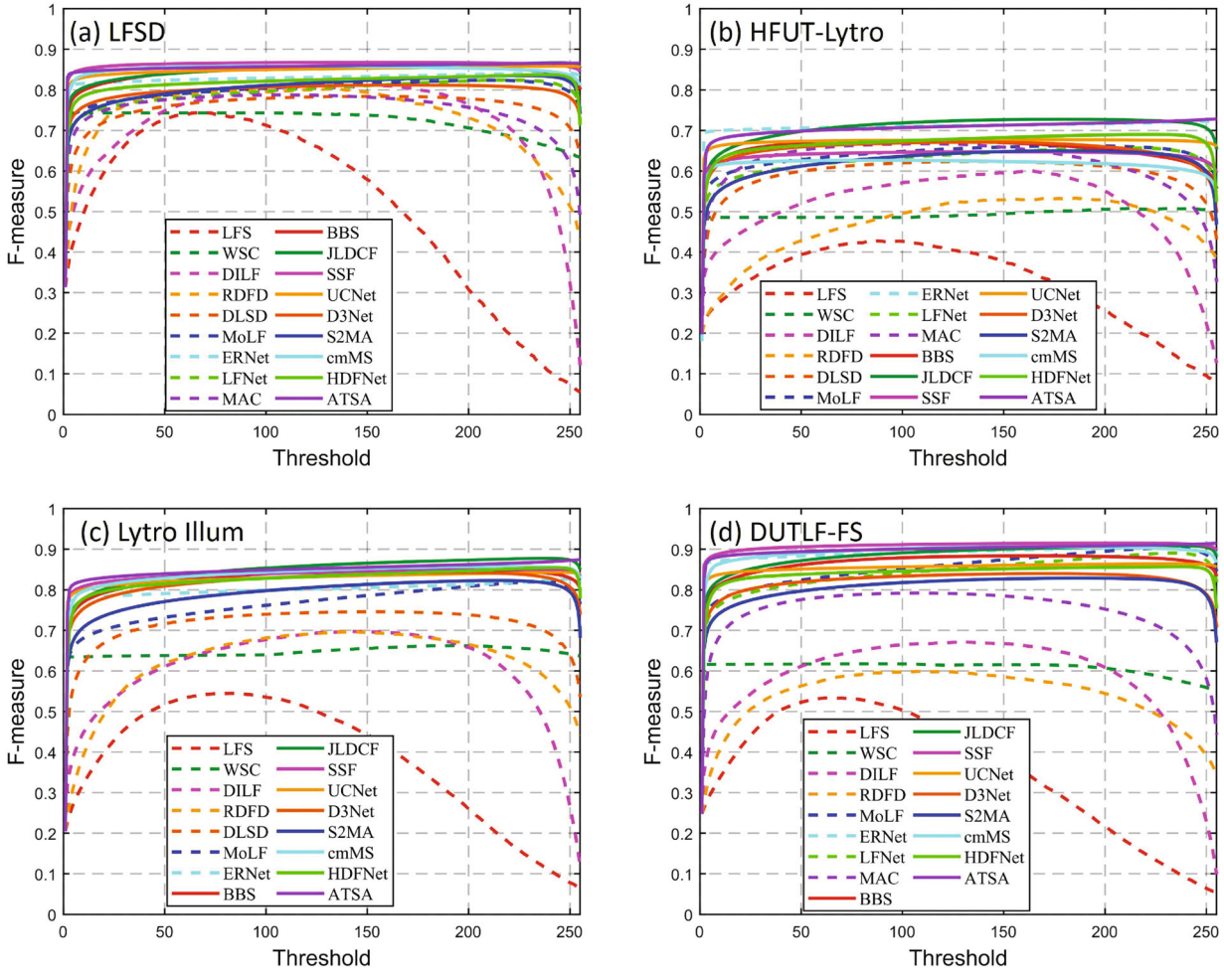


Fig. 13 F-measure curves for four datasets: (a) LFSD [5], (b) HFUT-Lytro [53], (c) Lytro Illum [59], and (d) DUTLF-FS [57], for nine light field SOD models: LFS [5], WSC [50], DILF [30], RDFD [54], DLSD [45], MoLF [31], ERNet [32], LFNet [58], MAC [59], and nine SOTA RGB-D based SOD models: BBS [35], JLDCF [33, 34], SSF [61], UCNet [62], D3Net [63], S2MA [64], cmMS [65], HDFNet [66], and ATSA [36]. *Solid lines* and *dashed lines* represent F-measure curves for *RGB-D based SOD models* and *light field SOD models*, respectively.

[33], while ERNet [32] was trained only on about 1000 light fields. Thus, the former is more likely to have better generalization.

However, we can still hardly deny the potential of light fields to boost the performance of SOD, as recently RGB-D SOD has been much more active, with many new competitive models proposed [49], than light field SOD. Furthermore, the performance of ERNet and MoLF is only slightly lower than that of the RGB-D models on the benchmark datasets, which further implies the effectiveness of light fields for SOD [48]. We believe that there is still considerable room for improving light field SOD, because light fields can provide more information than paired RGB and depth images.

Furthermore, in order to eliminate training discrepancies, we conducted experiments by retraining these RGB-D models on a unified training set, namely the training set of DUTLF-FS that contains 1000 scenarios. We also retrained ERNet to remove its extra HFUT-Lytro training data as shown in Table 1. Comparative results are given in Table 5, where all models generally incur certain performance degeneration. Interestingly, after retraining, SSF* can no longer outperform ERNet*, while ATSA* becomes inferior to ERNet* on LFSD and DUTLF-FS. Only JLDCF* and HDFNet* are consistently superior to ERNet* by a noticeable margin.

3.3.5 Accuracy across datasets

It is clearly shown in Table 4 and Fig. 12 that the models tested perform differently on different datasets. Generally, the models achieve better results on LFSD than on the other three datasets, indicating that LFSD is the easiest dataset for light field

Table 5 Quantitative measures: S-measure (S_α) [80], max F-measure ($F_\beta^{\max}$), max E-measure ($E_\phi^{\max}$), and MAE (M) [79] of one retrained light field SOD model (ERNet [32]) and seven retrained RGB-D based SOD models (i.e., BBS [35], SSF [61], ATSA [36], S2MA [64], D3Net [63], HDFNet [66], and JLDCF [33]). Results of original models are taken from Table 4, and the retrained models are marked by $*$. The best results of retrained models are highlighted in bold. ↑/↓ denotes that a larger/smaller value is better

Model	LFSD [5]				HFUT-Lytro [53]				Lytro Illum [59]				DUTLF-FS [57]			
	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$E_\phi^{\max}\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$E_\phi^{\max}\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$E_\phi^{\max}\uparrow$	$M\downarrow$	$S_\alpha\uparrow$	$F_\beta^{\max}\uparrow$	$E_\phi^{\max}\uparrow$	$M\downarrow$
BBS [35]	0.864	0.858	0.900	0.072	0.751	0.676	0.801	0.089	0.879	0.850	0.913	0.047	0.894	0.884	0.923	0.054
SSF [61]	0.859	0.868	0.901	0.067	0.725	0.647	0.778	0.100	0.872	0.850	0.913	0.044	0.908	0.915	0.946	0.036
ATSA [36]	0.858	0.866	0.902	0.068	0.772	0.729	0.833	0.084	0.883	0.875	0.929	0.041	0.905	0.915	0.943	0.039
ERNet [32]	0.831	0.842	0.884	0.083	0.778	0.722	0.841	0.082	0.843	0.827	0.911	0.056	0.899	0.908	0.949	0.039
S2MA [64]	0.837	0.835	0.873	0.094	0.729	0.650	0.777	0.112	0.853	0.823	0.895	0.063	0.845	0.829	0.883	0.079
D3Net [63]	0.825	0.812	0.863	0.095	0.749	0.671	0.797	0.091	0.869	0.843	0.909	0.050	0.852	0.840	0.891	0.069
HDFNet [66]	0.846	0.837	0.879	0.086	0.763	0.690	0.801	0.095	0.873	0.855	0.913	0.051	0.868	0.857	0.898	0.065
JLDCF [33]	0.862	0.867	0.902	0.070	0.789	0.727	0.844	0.075	0.890	0.878	0.931	0.042	0.905	0.908	0.943	0.043
BBS* [35]	0.739	0.738	0.812	0.123	0.708	0.622	0.773	0.102	0.825	0.788	0.878	0.065	0.873	0.870	0.919	0.051
SSF* [61]	0.790	0.793	0.861	0.097	0.687	0.612	0.781	0.099	0.833	0.799	0.886	0.059	0.881	0.889	0.930	0.050
ATSA* [36]	0.816	0.823	0.873	0.087	0.727	0.673	0.805	0.094	0.844	0.822	0.905	0.054	0.880	0.892	0.936	0.045
ERNet* [32]	0.822	0.825	0.885	0.085	0.707	0.632	0.766	0.117	0.840	0.810	0.900	0.059	0.898	0.903	0.946	0.040
S2MA* [64]	0.827	0.829	0.873	0.086	0.672	0.572	0.735	0.120	0.839	0.802	0.885	0.060	0.894	0.893	0.934	0.046
D3Net* [63]	0.827	0.821	0.877	0.086	0.720	0.645	0.801	0.092	0.859	0.835	0.906	0.051	0.906	0.911	0.947	0.039
HDFNet* [66]	0.849	0.850	0.891	0.073	0.747	0.673	0.801	**0.085**	0.874	0.854	0.915	0.045	0.922	**0.931**	0.955	**0.030**
JLDCF* [33]	**0.850**	**0.860**	**0.900**	**0.071**	**0.755**	**0.694**	**0.823**	0.086	**0.877**	**0.855**	**0.919**	**0.042**	**0.924**	**0.931**	**0.958**	**0.030**

SOD, on which the traditional model DILF can even outperform some deep models like DLSD and MAC. In contrast, HFUT-Lytro, Lytro Illum, and DUTLF-FS are more challenging. Note that MoLF, ERNet, ATSA work well on DUTLF-FS, probably because they were trained on DUTLF-FS's training set or training data (see Table 1). Besides, as mentioned in Section 2.3, HFUT-Lytro has many small salient objects, with multiple objects per image. The reduced performance of these models on this dataset tells that detecting small/multiple salient objects is still very challenging for existing schemes, both RGB-D based models and light field models. This makes HFUT-Lytro the most challenging among existing light field datasets.

3.3.6 Result visualization

Figure 14 visualizes some sample results from five light field models, including two traditional methods, LFS and DILF, three deep learning-based models, DLSD, MoLF, and ERNet, and three latest RGB-D based models, JLDCF, BBS, and ATSA. The top two rows in Fig. 14 show easy cases while the third to fifth rows show cases with complex backgrounds or sophisticated boundaries. The last row gives an example with low color contrast between foreground and background. As can be seen, RGB-D models perform comparably to or even better than light field models, which indicates that this field is still insufficiently studied. Figure 15 further shows several scenarios with small and multiple salient objects, where the first three rows show cases with multiple salient objects and others show cases with small objects. Both RGB-D based and light field models are more likely to make erroneous detections in such cases, confirming the poor abilities of existing techniques to handle small or multiple objects.

4 Challenges and open directions

This section highlights several future research directions for light field SOD and outlines several open issues.

4.1 Dataset collection and unification

As demonstrated in Section 2.3, existing light field datasets are limited in scale and have non-uniform data representations, making it somewhat difficult to evaluate different models and generalize deep networks. This non-uniformity issue is particularly severe for light field SOD because of its diverse data representations and high dependency on special acquisition hardware, unlike other SOD tasks such as RGB-D SOD [33, 36, 61] and video SOD [86, 87]. Therefore, developing large-scale and unified datasets is essential for future research. We urge researchers to take this issue into consideration when constructing new datasets. Moreover, collecting complete data forms, including raw data, focal stacks, multi-view images, depth maps, and micro-lens image arrays,

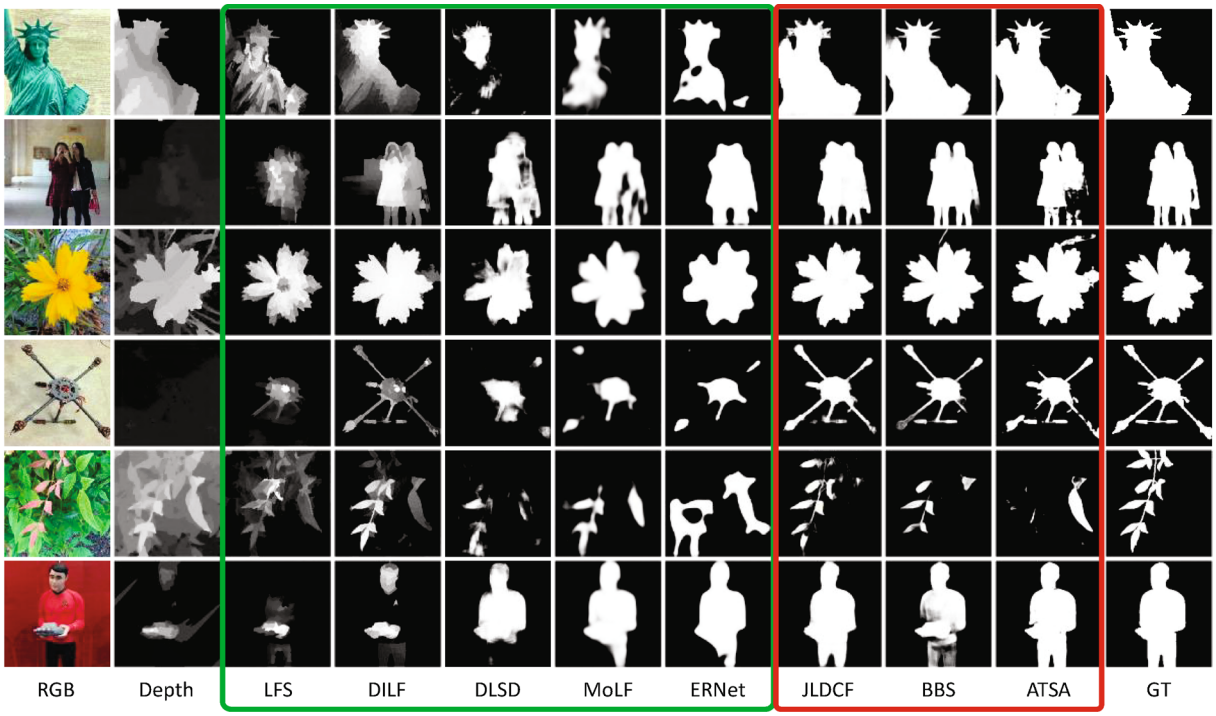


Fig. 14 Visual comparison of five light field SOD methods (green box): LFS [5], DILF [30], DLSD [45], MoLF [31], and ERNet [32], and three SOTA RGB-D based SOD models (red box): JLDCF [33, 34], BBS [35], and ATSA [36].

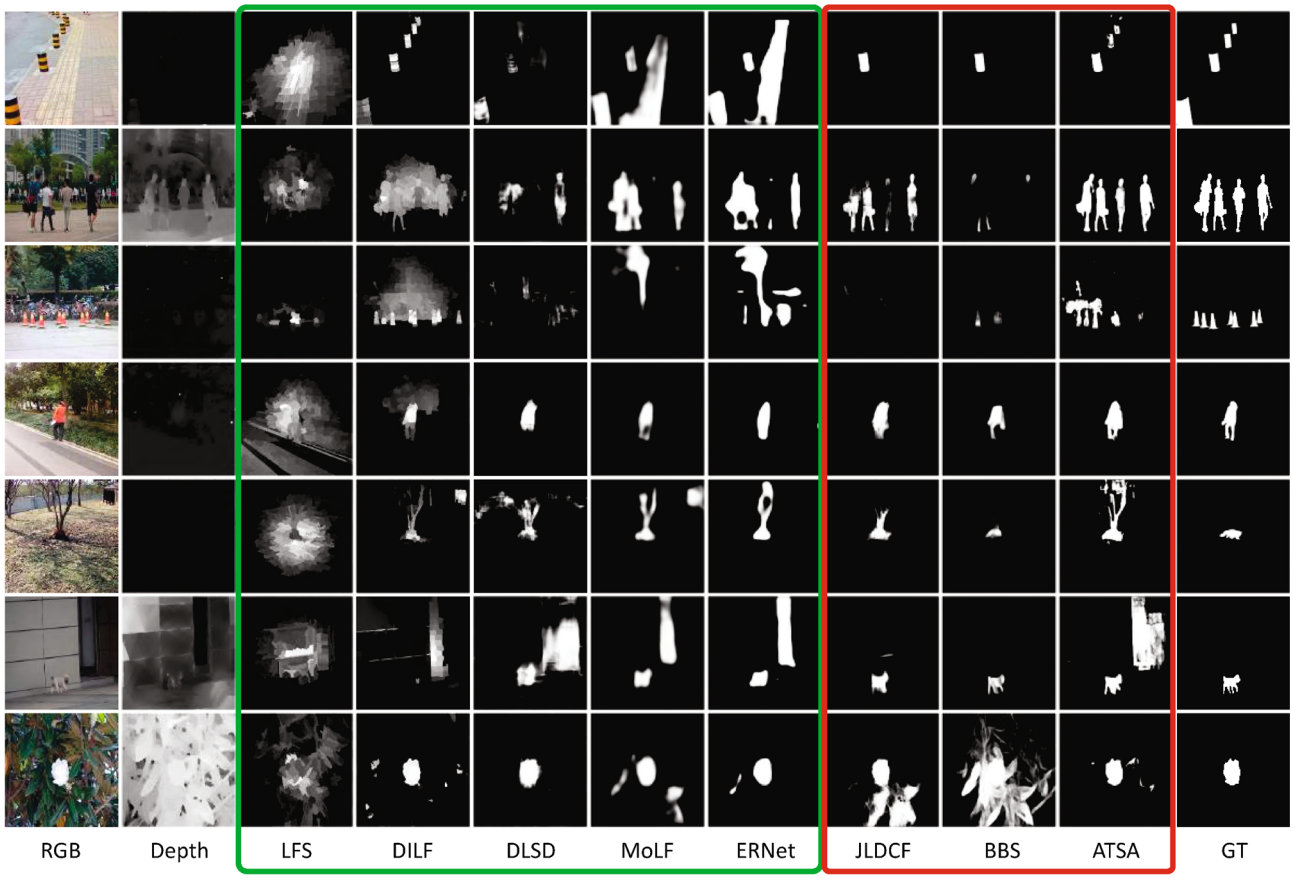


Fig. 15 Comparison of five light field SOD models (green box): LFS [5], DILF [30], DLSD [45], MoLF [31], and ERNet [32], and three SOTA RGB-D based SOD models (red box): JLDCF [33, 34], BBS [35], and ATSA [36], when detecting small and multiple objects.

would definitely facilitate and advance research on this topic. However, we also note that there is a challenge in data storage and transmission, since raw light field data is quite large in size (e.g., the 640 light fields of Lytro Illum occupy 32.8 gigabytes), not to mention a large-scale dataset. The scale of the dataset makes it a bit difficult to spread. In this case, it will still be great if a subset of any data form is available for the public.

4.2 Developing light field SOD

As noted, there are currently fewer studies on SOD for light fields than for other tasks in the saliency community. Thus, this field is still under-explored. From the benchmarking results in Section 3.3, it can be observed that the SOTA performance is still far from satisfactory, especially on the HFUT-Lytro dataset. There is considerable room for further improvement of light field SOD algorithms and models. In addition, we note that only seven deep learning-based models appeared between 2019 and 2020. We attribute such a scarcity of light field SOD research to the aforementioned data problems, as well as the lack of a comprehensive survey of existing methods and datasets for this topic.

4.3 Multi-view images and micro-lens image arrays

Most existing models work with focal stacks and depth maps, as shown in Table 1, while multi-view images and micro-lens image arrays are two other types of light field data representations that are rarely considered (in only five models). The benchmarking results in Section 3.3 show that the latter do not perform as well as models utilizing other data forms, so the use of these two data forms has not yet been fully explored. Thus, more work on light field SOD models is needed to explore the effectiveness of multi-view images and micro-lens image arrays. Alternatively, these two data representations themselves may be less informative than focal stacks and depth maps: Scene depth information may be more implicitly conveyed. This may make it difficult to find effective mappings and mine underlying rules using deep neural networks, especially when the training data is sparse. It would be interesting to compare the effectiveness and redundancy of saliency detection using different data representations.

4.4 Incorporating high-quality depth estimation

It has been shown that accurate depth maps are conducive to discovering salient objects from complex backgrounds. Unfortunately, the quality of depth maps varies greatly in several existing datasets, since depth estimation from light fields is a challenging task [38–41, 43, 44]. The challenge stems from the fact that although the light fields can be used to synthesize images focused at any depth through digital refocusing technology, the depth distribution of each scene point is unknown. Besides, it is necessary to determine whether the image area is in focus, which itself is a difficult issue [88, 89]. Imperfect depth maps often negatively impact the detection accuracy of models using depth maps. Therefore, incorporating high-quality depth estimation algorithms from light fields is likely to be beneficial.

4.5 Edge-aware light field SOD

Accurate object boundaries are essential for high-quality saliency maps, as SOD is a pixel-wise segmentation task [3]. In the RGB SOD field, edge-aware SOD models are drawing increasing research attention [28, 83, 84]. Currently, as shown in our experimental results, existing deep light field SOD models rarely consider this issue, resulting in saliency maps with coarse boundaries and edges. Thus, edge-aware light field SOD should be a future research direction.

4.6 Acquisition technology and hardware

The first generation light field camera, Lytro, was invented in 2011, while its successor, Lytro Illum, was introduced in 2014. The latter is more powerful but is much larger than the former, and is also much more expensive. However, in general, the development of light field acquisition technology and hardware has been slower than that of, e.g., computers and mobile phones. Since 2014, there have been few commercial light field cameras. There is an urgent need for the development of acquisition and hardware technology for light field photography. Currently, light field cameras are far from replacing traditional RGB cameras in terms of image quality, price, and portability. If in future light field cameras were to become affordable and small, they could easily be integrated into mobile phones, allowing everyone to try light field photography in daily life. This would provide a vast increase in user data and post-

processing application requirements, paving the way for significant improvements in light field SOD.

4.7 Supervision strategies

Existing deep light field models learn to segment salient objects in a fully supervised manner, which requires sufficient annotated training data. Unfortunately, the size of the existing datasets is limited: DUTLF-FS and DUTLF-MV provide 1000 and 1100 samples for training, respectively, while other datasets contain fewer than 640 light fields. On one hand, a small amount of training data limits the generalizability of models. On the other hand, acquiring a large amount of annotated data requires extreme manual effort for data collection and labelling. Recently, weakly- and semi-supervised learning strategies have attracted extensive research attention, largely reducing the annotation effort. Being data-friendly, they have been introduced into RGB SOD, and some encouraging attempts [8, 90, 91] have been made. Thus, one future direction is to extend these supervision strategies to light field SOD, to overcome the shortages of training data. Additionally, several works [92, 93] have shown that pre-training models in a self-supervised manner can effectively improve performance, which could also be introduced to light field SOD in future.

4.8 Linking RGB-D SOD to light field SOD

There is a close connection between light field SOD and RGB-D SOD, since both tasks explore scene depth information for saliency detection, while depth information can be derived from light field data using a variety of techniques. This is why RGB-D SOD can be regarded as a solution to the degradation of light field SOD. As shown in Table 4, applying RGB-D SOD models to light fields is straightforward, whereas we believe the reverse could also be possible. For example, intuitively, reconstructing light field data such as focal stacks or multi-view images from a pair of RGB and depth images is possible [45]. If this bridge is realized, mutual transfer between the models of these two fields becomes feasible, and then light field models can be applied to RGB-D data. Such a link would be an interesting issue to explore in the near future.

4.9 Other potential directions

Inspired by recent advances in the saliency community, there are several other potential directions for future research. For example, high-resolution salient object detection [94] aims to deal with salient object segmentation in high-resolution images, and achieving high-resolution details could be considered in light field SOD. Besides, while existing light field datasets are labelled at an object-level, instance-level annotation and detection, which aim to separate individual objects [95–99], could also be introduced into this field. There are many instance-sensitive application scenarios, e.g., image captioning [100], and multi-label image recognition [101], as well as various weakly supervised/unsupervised learning scenarios [102, 103]. Recent work has attempted to address weakly-supervised salient instance detection [104]. Similarly, more effort could be spent on instance-level ground-truth annotation and designing instance-level light field SOD models. Furthermore, eye-fixation prediction [3, 105, 106] is another subfield of saliency detection. So far, there has been no research on eye-fixation prediction using light field data. As abundant natural scene information is provided by the light field, we hope that the various data representations of the light field could provide useful cues to help eliminate ambiguous eye-fixation. Lastly, light field data could benefit other tasks closely related to SOD, such as camouflaged object detection (COD) [107] and transparent object segmentation [108], where objects often borrow texture from their background and have similar appearances to their surroundings.

Finally, there is an unanswered question remaining: How can light field information be more beneficial to SOD than depth information? Depth information can be derived from and is a subset of light field data. Different forms of light field data, e.g., focal stacks and multi-view images, somewhat imply depth information, indicating that existing models may implicitly leverage such depth information. So what is the difference between using depth in an explicit way (like RGB-D SOD models) and in an implicit way? This is an interesting question, but unfortunately, since the problem of light field SOD was proposed in 2014, no study has shown any direct answer or evidence. This is worthy of further investigation and understanding in future.

5 Conclusions

We have provided the first comprehensive review and benchmark for light field SOD, reviewing and

discussing existing studies and related datasets. We have benchmarked representative light field SOD models and compared them to several cutting-edge RGB-D SOD models both qualitatively and quantitatively. As existing light field datasets are somewhat inconsistent in data representations, we have generated supplemental data for existing datasets, making them complete and uniform. Moreover, we have discussed several potential directions for future research and outlined some open issues. Although progress has been made over the past several years, there are still only seven deep learning-based works on this topic, leaving significant room to design more powerful network architectures incorporating effective modules like edge-aware designs and top–down refinement, to improve SOD performance. We hope this survey will serve as a catalyst to advance this area and promote interesting work in future.

Acknowledgements

Keren Fu was supported by the National Natural Science Foundation of China (Nos. 62176169 and 61703077) and SCU-Luzhou Municipal People's Government Strategic Cooperation Project (No. 2020CDLZ-10). Tao Zhou was supported by the National Natural Science Foundation of China (No. 62172228). Qijun Zhao was supported by the National Natural Science Foundation of China (No. 61773270).

Declaration of competing interest

The authors have no competing interests to declare that are relevant to the content of this article.

References

[1] Wang, S. Z.; Liao, W. J.; Surman, P.; Tu, Z. G.; Zheng, Y. J.; Yuan, J. S. Salience guided depth calibration for perceptually optimized compressive light field 3D display. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2031–2040, 2018.

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

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

[4] Borji, A.; Cheng, M. M.; Hou, Q. B.; Jiang, H. Z.; Li, J. Salient object detection: A survey. *Computational Visual Media* Vol. 5, No. 2, 117–150, 2019.

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

[6] Li, N. Y.; Ye, J. W.; Ji, Y.; Ling, H. B.; Yu, J. Y. Saliency detection on light field. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 39, No. 8, 1605–1616, 2017.

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

[8] Zhang, D.; Meng, D.; Zhao, L.; Han, J. Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. In: Proceedings of the 25th International Joint Conference on Artificial Intelligence, 3538–3544, 2016.

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

[10] Moosmann, F.; Larlus, D.; Jurie, F. Learning saliency maps for object categorization. In: Proceedings of the ECCV'06 Workshop on the Representation and Use of Prior Knowledge in Vision, 2006.

[11] Cheng, M. M.; Liu, Y.; Lin, W. Y.; Zhang, Z. M.; Rosin, P. L.; Torr, P. H. S. BING: Binarized normed gradients for objectness estimation at 300fps. *Computational Visual Media* Vol. 5, No. 1, 3–20, 2019.

[12] Wei, Y. C.; Feng, J. S.; Liang, X. D.; Cheng, M. M.; Zhao, Y.; Yan, S. C. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 6488–6496, 2017.

[13] Wei, Y. C.; Liang, X. D.; Chen, Y. P.; Shen, X. H.; Cheng, M. M.; Feng, J. S.; Zhao, Y.; Yan, S. STC: A simple to complex framework for weakly-supervised semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 39, No. 11, 2314–2320, 2017.

[14] Wang, X.; You, S. D.; Li, X.; Ma, H. M. Weakly-supervised semantic segmentation by iteratively mining common object features. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1354–1362, 2018.

[15] Wang, W.; Shen, J.; Yang, R.; Porikli, F. Saliency-aware video object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 40, No. 1, 20–33, 2018.

[16] Song, H.; Wang, W.; Zhao, S.; Shen, J.; Lam, K.-M. Pyramid dilated deeper ConvLSTM for video salient object detection. In: *Computer Vision – ECCV 2018. ECCV 2018. Lecture Notes in Computer Science, Vol. 11215*. Ferrari, V.; Hebert, M.; Sminchisescu, C.; Weiss, Y. Eds. Springer Cham, 744–760, 2018.

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

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

[19] Ma, Y. F.; Lu, L.; Zhang, H. J.; Li, M. J. A user attention model for video summarization. In: Proceedings of the 10th ACM International Conference on Multimedia, 533–542, 2002.

[20] Ouerhani, N.; Bracamonte, J.; Hugli, H.; Ansorge, M.; Pellandini, F. Adaptive color image compression based on visual attention. In: Proceedings of the 11th International Conference on Image Analysis and Processing, 416–421, 2001.

[21] Han, J. G.; Pauwels, E. J.; de Zeeuw, P. Fast saliency-aware multi-modality image fusion. *Neurocomputing* Vol. 111, 70–80, 2013.

[22] Jin, S.; Ling, H. B. Scale and object aware image retargeting for thumbnail browsing. In: Proceedings of the International Conference on Computer Vision, 1511–1518, 2011.

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

[24] Borji, A.; Itti, L. Defending Yarbus: Eye movements reveal observers' task. *Journal of Vision* Vol. 14, No. 3, 29, 2014.

[25] Fu, K. R.; Zhao, Q. J.; Yu-Hua Gu, I.; Yang, J. Deepside: A general deep framework for salient object detection. *Neurocomputing* Vol. 356, 69–82, 2019.

[26] Wang, W. G.; Shen, J. B.; Shao, L.; Porikli, F. Correspondence driven saliency transfer. *IEEE Transactions on Image Processing* Vol. 25, No. 11, 5025–5034, 2016.

[27] Zhang, P. P.; Wang, D.; Lu, H. C.; Wang, H. Y.; Ruan, X. Amulet: Aggregating multi-level convolutional features for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision, 202–211, 2017.

[28] Feng, M. Y.; Lu, H. C.; Ding, E. R. Attentive feedback network for boundary-aware salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 1623–1632, 2019.

[29] Zhang, P. P.; Wang, D.; Lu, H. C.; Wang, H. Y.; Yin, B. C. Learning uncertain convolutional features for accurate saliency detection. In: Proceedings of the IEEE International Conference on Computer Vision, 212–221, 2017.

[30] Zhang, J.; Wang, M.; Gao, J.; Wang, Y.; Zhang, X.; Wu, X. Saliency detection with a deeper investigation of light field. In: Proceedings of the 24th International Conference on Artificial Intelligence, 2212–2218, 2015.

[31] Zhang, M.; Li, J.; Ji, W.; Piao, Y.; Lu, H. Memory-oriented decoder for light field salient object detection. In: Proceedings of the 33rd Conference on Neural Information Processing Systems, 898–908, 2019.

[32] Piao, Y. R.; Rong, Z. K.; Zhang, M.; Lu, H. C. Exploit and replace: An asymmetrical two-stream architecture for versatile light field saliency detection. *Proceedings of the AAAI Conference on Artificial Intelligence* Vol. 34, No. 7, 11865–11873, 2020.

[33] Fu, K. R.; Fan, D. P.; Ji, G. P.; Zhao, Q. J. JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3049–3059, 2020.

[34] Fu, K. R.; Fan, D. P.; Ji, G. P.; Zhao, Q. J.; Shen, J. B.; Zhu, C. Siamese network for RGB-D salient object detection and beyond. *IEEE Transactions on Pattern Analysis and Machine Intelligence* doi:10.1109/TPAMI.2021.3073689, 2021.

[35] Fan, D. P.; Zhai, Y.; Borji, A.; Yang, J.; Shao, L. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12357*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 275–292, 2020.

[36] Zhang, M.; Fei, S. X.; Liu, J.; Xu, S.; Piao, Y. R.; Lu, H. C. Asymmetric two-stream architecture for accurate RGB-D saliency detection. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12373*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 374–390, 2020.

[37] Gershun, A. The light field. *Studies in Applied Mathematics* Vol. 18, Nos. 1–4, 51–151, 1939.

[38] Jeon, H. G.; Park, J.; Choe, G.; Park, J.; Bok, Y.; Tai, Y. W.; Kweon, I.-S. Accurate depth map estimation from a lenslet light field camera. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1547–1555, 2015.

[39] Tao, M. W.; Hadap, S.; Malik, J.; Ramamoorthi, R. Depth from combining defocus and correspondence using light-field cameras. In: Proceedings of the IEEE International Conference on Computer Vision, 673–680, 2013.

[40] Tao, M. W.; Srinivasan, P. P.; Malik, J.; Rusinkiewicz, S.; Ramamoorthi, R. Depth from shading, defocus, and correspondence using light-field angular coherence. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 1940–1948, 2015.

[41] Wang, T. C.; Efros, A. A.; Ramamoorthi, R. Occlusion-aware depth estimation using light-field cameras. In: Proceedings of the IEEE International Conference on Computer Vision, 3487–3495, 2015.

[42] Ng, R.; Levoy, M.; Brédif, M.; Duval, G.; Horowitz, M.; Hanrahan, P. Light field photography with a hand-held plenopic camera. Stanford Tech Report CTSR 2005-02, 2005.

[43] Piao, Y.; Zhang, Y.; Zhang, M.; Ji, X. Dynamic fusion network for light field depth estimation. *arXiv preprint* arXiv:2104.05969, 2021.

[44] Piao, Y.; Ji, X.; Zhang, M.; Zhang, Y. Learning multi-modal information for robust light field depth estimation. *arXiv preprint* arXiv:2104.05971, 2021.

[45] Piao, Y.; Rong, Z.; Zhang, M.; Li, X.; Lu, H. Deep light-field-driven saliency detection from a single view. In: Proceedings of the 28th International Joint Conference on Artificial Intelligence, 904–911, 2019.

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

[47] Buehler, C.; Bosse, M.; McMillan, L.; Gortler, S.; Cohen, M. Unstructured lumigraph rendering. In: Proceedings of the 28th Annual Conference on Computer Graphics and Interactive Techniques, 425–432, 2001.

[48] Zhang, X. D.; Wang, Y.; Zhang, J.; Hu, L. M.; Wang, M. Light field saliency vs. 2D saliency: A comparative study. *Neurocomputing* Vol. 166, 389–396, 2015.

[49] Zhou, T.; Fan, D. P.; Cheng, M. M.; Shen, J. B.; Shao, L. RGB-D salient object detection: A survey. *Computational Visual Media* Vol. 7, No. 1, 37–69, 2021.

[50] Li, N. Y.; Sun, B. L.; Yu, J. Y. A weighted sparse coding framework for saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 5216–5223, 2015.

[51] Sheng, H.; Zhang, S.; Liu, X. Y.; Xiong, Z. Relative location for light field saliency detection. In: Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing, 1631–1635, 2016.

[52] Wang, A. Z.; Wang, M. H.; Li, X. Y.; Mi, Z. T.; Zhou, H. A two-stage Bayesian integration framework for salient object detection on light field. *Neural Processing Letters* Vol. 46, No. 3, 1083–1094, 2017.

[53] Zhang, J.; Wang, M.; Lin, L.; Yang, X.; Gao, J.; Rui, Y. Saliency detection on light field. *ACM Transactions on Multimedia Computing, Communications, and Applications* Vol. 13, No. 3, Article No. 32, 2017.

[54] Wang, X.; Dong, Y. Y.; Zhang, Q.; Wang, Q. Region-based depth feature descriptor for saliency detection on light field. *Multimedia Tools and Applications* Vol. 80, No. 11, 16329–16346, 2021.

[55] Piao, Y. R.; Li, X.; Zhang, M.; Yu, J. Y.; Lu, H. C. Saliency detection via depth-induced cellular automata on light field. *IEEE Transactions on Image Processing* Vol. 29, 1879–1889, 2020.

[56] Wang, H. Q.; Yan, B.; Wang, X. Z.; Zhang, Y. B.; Yang, Y. Accurate saliency detection based on depth feature of 3D images. *Multimedia Tools and Applications* Vol. 77, No. 12, 14655–14672, 2018.

[57] Wang, T. T.; Piao, Y. R.; Lu, H. C.; Li, X.; Zhang, L. H. Deep learning for light field saliency detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 8837–8847, 2019.

[58] Zhang, M.; Ji, W.; Piao, Y. R.; Li, J. J.; Zhang, Y.; Xu, S.; Lu, H. LFNet: Light field fusion network for salient object detection. *IEEE Transactions on Image Processing* Vol. 29, 6276–6287, 2020.

[59] Zhang, J.; Liu, Y. M.; Zhang, S. P.; Poppe, R.; Wang, M. Light field saliency detection with deep convolutional networks. *IEEE Transactions on Image Processing* Vol. 29, 4421–4434, 2020.

[60] Zhang, Q. D.; Wang, S. Q.; Wang, X.; Sun, Z. H.; Kwong, S.; Jiang, J. M. A multi-task collaborative network for light field salient object detection. *IEEE Transactions on Circuits and Systems for Video Technology* Vol. 31, No. 5, 1849–1861, 2021.

[61] Zhang, M.; Ren, W. S.; Piao, Y. R.; Rong, Z. K.; Lu, H. C. Select, supplement and focus for RGB-D saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3469–3478, 2020.

[62] Zhang, J.; Fan, D. P.; Dai, Y. C.; Anwar, S.; Saleh, F. S.; Zhang, T.; Barnes, N. UC-net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders. In: Proceedings of the

IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8579–8588, 2020.

[63] Fan, D. P.; Lin, Z.; Zhang, Z.; Zhu, M. L.; Cheng, M. M. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems* Vol. 32, No. 5, 2075–2089, 2021.

[64] Liu, N.; Zhang, N.; Han, J. W. Learning selective self-mutual attention for RGB-D saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 13753–13762, 2020.

[65] Li, C. Y.; Cong, R. M.; Piao, Y. R.; Xu, Q. Q.; Loy, C. C. RGB-D salient object detection with cross-modality modulation and selection. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12353*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 225–241, 2020.

[66] Pang, Y.; Zhang, L.; Zhao, X.; Lu, H. Hierarchical dynamic filtering network for RGB-D salient object detection. In: *Computer Vision – ECCV 2020. Lecture Notes in Computer Science, Vol. 12370*. Vedaldi, A.; Bischof, H.; Brox, T.; Frahm, J. M. Eds. Springer Cham, 235–252, 2020.

[67] Adelson, E.; Bergen, J. The Plenoptic function and the elements of early vision. In: *Computational Models of Visual Processing*. MIT Press, 3–20, 1991.

[68] Levoy, M.; Hanrahan, P. Light field rendering. In: Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques, 31–42, 1996.

[69] Agarwala, A.; Dontcheva, M.; Agrawala, M.; Drucker, S.; Colburn, A.; Curless, B.; Salesin, D.; Cohen, M. Interactive digital photomontage. *ACM Transactions on Graphics* Vol. 23, No. 3, 294–302, 2004.

[70] Kuthirummal, S.; Nagahara, H.; Zhou, C. Y.; Nayar, S. K. Flexible depth of field photography. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 33, No. 1, 58–71, 2011.

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

[72] He, K. M.; Sun, J.; Tang, X. O. Single image haze removal using dark channel prior. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 33, No. 12, 2341–2353, 2011.

[73] Shelhamer, E.; Long, J.; Darrell, T. Fully convolutional networks for semantic segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 39, No. 4, 640–651, 2017.

[74] Shi, X.; Chen, Z.; Wang, H.; Yeung, D.; Wong, W.; Woo, W. Convolutional LSTM network: A machine learning approach for precipitation nowcasting. In: Proceedings of the 28th International Conference on Neural Information Processing Systems, 802–810, 2015.

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

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

[77] Wang, L. J.; Lu, H. C.; Wang, Y. F.; Feng, M. Y.; Wang, D.; Yin, B. C.; Ruan, X. Learning to detect salient objects with image-level supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3796–3805, 2017.

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

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

[80] Fan, D. P.; Cheng, M. M.; Liu, Y.; Li, T.; Borji, A. Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE International Conference on Computer Vision, 4558–4567, 2017.

[81] Zhao, J. X.; Cao, Y.; Fan, D. P.; Cheng, M. M.; Li, X. Y.; Zhang, L. Contrast prior and fluid pyramid integration for RGBD salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 3922–3931, 2019.

[82] Fan, D. P.; Gong, C.; Cao, Y.; Ren, B.; Cheng, M. M.; Borji, A. Enhanced-alignment measure for binary foreground map evaluation. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, 698–704, 2018.

[83] Wu, Z.; Su, L.; Huang, Q. M. Stacked cross refinement network for edge-aware salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 7263–7272, 2019.

[84] Qin, X. B.; Zhang, Z. C.; Huang, C. Y.; Gao, C.; Dehghan, M.; Jagersand, M. BASNet: Boundary-aware salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 7471–7481, 2019.

[85] Li, G. Y.; Liu, Z.; Ling, H. B. ICNet: Information conversion network for RGB-D based salient object detection. *IEEE Transactions on Image Processing* Vol. 29, 4873–4884, 2020.

[86] Tsiami, A.; Koutras, P.; Maragos, P. STAViS: Spatio-temporal AudioVisual saliency network. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4765–4775, 2020.

[87] Fan, D. P.; Wang, W. G.; Cheng, M. M.; Shen, J. B. Shifting more attention to video salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8546–8556, 2019.

[88] Zhao, W. D.; Zhao, F.; Wang, D.; Lu, H. C. Defocus blur detection via multi-stream bottom-top-bottom network. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 42, No. 8, 1884–1897, 2020.

[89] Park, J.; Tai, Y. W.; Cho, D.; Kweon, I. S. A unified approach of multi-scale deep and hand-crafted features for defocus estimation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2760–2769, 2017.

[90] Zeng, Y.; Zhuge, Y. Z.; Lu, H. C.; Zhang, L. H.; Qian, M. Y.; Yu, Y. Z. Multi-source weak supervision for saliency detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 6067–6076, 2019.

[91] Qian, M. Y.; Qi, J. Q.; Zhang, L. H.; Feng, M. Y.; Lu, H. C. Language-aware weak supervision for salient object detection. *Pattern Recognition* Vol. 96, 106955, 2019.

[92] Chen, T. L.; Liu, S. J.; Chang, S. Y.; Cheng, Y.; Amini, L.; Wang, Z. Y. Adversarial robustness: From self-supervised pre-training to fine-tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 696–705, 2020.

[93] Dai, A.; Diller, C.; Niessner, M. SG-NN: Sparse generative neural networks for self-supervised scene completion of RGB-D scans. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 846–855, 2020.

[94] Zeng, Y.; Zhang, P. P.; Lin, Z.; Zhang, J. M.; Lu, H. C. Towards high-resolution salient object detection. In: Proceedings of the IEEE/CVF International Conference on Computer Vision, 7233–7242, 2019.

[95] Cai, Z. W.; Vasconcelos, N. Cascade R-CNN: High quality object detection and instance segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 43, No. 5, 1483–1498, 2021.

[96] Chen, K.; Pang, J. M.; Wang, J. Q.; Xiong, Y.; Li, X. X.; Sun, S. Y.; Feng, W.; Liu, Z.; Shi, J.; Ouyang, W.; et al. Hybrid task cascade for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 4969–4978, 2019.

[97] Liu, S.; Qi, L.; Qin, H. F.; Shi, J. P.; Jia, J. Y. Path aggregation network for instance segmentation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 8759–8768, 2018.

[98] Li, G.; Xie, Y.; Lin, L.; Yu, Y. Instance-level salient object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 247–256, 2017.

[99] Fan, R. C.; Cheng, M. M.; Hou, Q. B.; Mu, T. J.; Wang, J. D.; Hu, S. M. S4Net: Single stage salient-instance segmentation. *Computational Visual Media* Vol. 6, No. 2, 191–204, 2020.

[100] Karpathy, A.; Li, F. F. Deep visual-semantic alignments for generating image descriptions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 39, No. 4, 664–676, 2017.

[101] Wei, Y.; Xia, W.; Huang, J.; Ni, B.; Dong, J.; Zhao, Y.; Yan, S. CNN: Single-label to multi-label. *arXiv preprint* arXiv:1406.5726, 2014.

[102] Chen, X. L.; Gupta, A. Webly supervised learning of convolutional networks. In: Proceedings of the IEEE International Conference on Computer Vision, 1431–1439, 2015.

[103] Lai, B. S.; Gong, X. J. Saliency guided dictionary learning for weakly-supervised image parsing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 3630–3639, 2016.

[104] Tian, X.; Xu, K.; Yang, X.; Yin, B.; Lau, R. Weakly-supervised salient instance detection. In: Proceedings of the Conference on British Machine Vision Conference, 2020.

[105] Borji, A. Saliency prediction in the deep learning era: Successes and limitations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* Vol. 43, No. 2, 679–700, 2021.

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

[107] Fan, D. P.; Ji, G. P.; Sun, G. L.; Cheng, M. M.; Shen, J. B.; Shao, L. Camouflaged object detection.

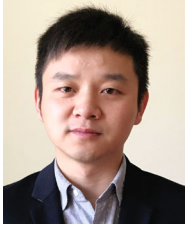
In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2774–2784, 2020.

[108] Xu, Y. C.; Nagahara, H.; Shimada, A.; Taniguchi, R. I. TransCut: Transparent object segmentation from a light-field image. In: Proceedings of the IEEE International Conference on Computer Vision, 3442–3450, 2015.

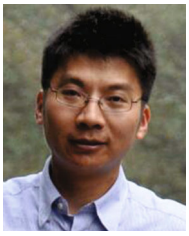
Keren Fu received his dual Ph.D. degrees from Shanghai Jiao Tong University, China, and Chalmers University of Technology, Sweden, under the joint supervision of Prof. Jie Yang and Prof. Irene Yu-Hua Gu. He is currently a research associate professor with the College of Computer Science, Sichuan University, China. His current research interests include visual computing, saliency analysis, and machine learning.

Yao Jiang is currently pursuing his master degree in the College of Computer Science, Sichuan University under the supervision of Dr. Keren Fu. His research interests include machine learning and computer vision.

Ge-Peng Ji received his master degree in communication and information systems from the School of Computer Science, Wuhan University, China. He is currently a research intern at the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates. His research interests lie in designing deep neural networks and applying deep learning to various fields of low-level vision, such as camouflaged/salient object detection, video salient object detection, and medical image segmentation.



Tao Zhou received his Ph.D. degree in pattern recognition and intelligent systems from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, in 2016. From 2016 to 2018, he was a postdoctoral fellow in the BRIC and IDEA Lab, University of North Carolina at Chapel Hill. From 2018 to 2020, he was a research scientist at the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi. He is currently a professor in the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include machine learning, computer vision, and medical image analysis.



Qijun Zhao is currently a professor in the College of Computer Science at Sichuan University. He is also a visiting professor in the School of Information Science and Technology at Tibet University. He obtained his B.Sc. and M.Sc. degrees in computer science both from Shanghai Jiao Tong University, and his Ph.D. degree in computer science from Hong Kong Polytechnic University. He worked as a post-doctoral research fellow in the Pattern Recognition and Image Processing Lab at Michigan State University from 2010 to 2012. His research is in the fields of pattern recognition, image processing, and computer vision.

Deng-Ping Fan received his Ph.D. degree from Nankai University in 2019. He was a research scientist in the Inception Institute of Artificial Intelligence (IIAI) during 2019–2021, and joined ETH Zürich as a postdoc researcher in 2021. He has published over 30 top journal and conference papers. His research interests include computer vision, deep learning, and saliency detection, especially human vision for co-salient object detection, RGB salient object detection, RGB-D salient object detection, and video salient object detection.