RGB-D Salient Object Detection: A Survey

Tao Zhou¹, Deng-Ping Fan¹(✉), Ming-Ming Cheng², Jianbing Shen¹, and Ling Shao¹

© The Author(s) 2020. This article is published with open access at Springerlink.com

Abstract Salient object detection (SOD), which simulates the human visual perception system to locate the most attractive object(s) in a scene, has been widely applied to various computer vision tasks. Now, with the advent of depth sensors, depth maps with affluent spatial information that can be beneficial in boosting the performance of SOD, can easily be captured. Although various RGB-D based SOD models with promising performance have been proposed over the past several years, an in-depth understanding of these models and challenges in this topic remains lacking. In this paper, we provide a comprehensive survey of RGB-D based SOD models from various perspectives, and review related benchmark datasets in detail. Further, considering that the light field can also provide depth maps, we review SOD models and popular benchmark datasets from this domain as well. Moreover, to investigate the SOD ability of existing models, we carry out a comprehensive evaluation, as well as attribute-based evaluation of several representative RGB-D based SOD models. Finally, we discuss several challenges and open directions of RGB-D based SOD for future research. All collected models, benchmark datasets, source code links, datasets constructed for attribute-based evaluation, and codes for evaluation will be made publicly available at https://github.com/taozh2017/RGBD-SODsurvey.

Keywords RGB-D based salient object detection, saliency detection, comprehensive evaluation, light field.

1 Introduction

Salient object detection (SOD) aims to locate the most visually prominent object(s) in a given scene [28]. SOD plays a key role in a range of real-world applications, such as stereo matching [95], image understanding [192], co-saliency detection [33], action recognition [110], video detection and segmentation [35, 118, 140, 141], semantic segmentation [115, 153], medical image segmentation [32, 37, 145], object tracking [49, 91], person re-identification [92, 175], camouflaged object detection [31], etc. Although significant progress has made in the SOD field over the past several years [50, 73, 73, 77, 119, 136, 142, 150, 151, 157, 168, 174, 176], there is still room for improvement under challenging factors such as complicated background or different lighting conditions.
in the scenes. One way to overcome these challenges is to employ depth maps, which provide complementary spatial information for RGB images and have become easier to capture due to the large availability of depth sensors (e.g., Microsoft Kinect).

Recently, RGB-D based SOD has gained increasing attention and various methods have been developed [3, 34]. Early RGB-D based SOD models tended to extract handcrafted features and then fuse RGB image and depth maps. For example, Lang et al. [65], the first work on RGB-D based SOD, utilized Gaussian mixture models to model the distribution of depth-induced saliency. Ciptadi et al. [16] extracted 3D layout and shape features from depth measurements. Besides, several methods [14, 14, 24, 111] measured depth contrast using the depth difference between different regions. In [99], a multi-contextual contrast model including local, global, and background contrast was developed to detect salient objects using depth maps. More importantly, however, this work also provided the first large-scale RGB-D dataset for SOD. Despite the effectiveness achieved by traditional methods using handcrafted features, they tend to suffer from a limited generalization ability for low-level features and lack the high-level reasoning required for complex scenes. To address these limitations, deep learning based RGB-D SOD methods [34] have been developed, showing improved performance. DF [109] was the first model to introduce deep learning technology into the RGB-D based SOD task. More recently, various deep learning-based models [9, 42, 71, 82, 101, 160, 173] have focused on exploiting effective multi-modal correlations and multi-scale/level information to boost SOD performance. To clearly describe the progress in the RGB-D based SOD field, we provide a brief chronology in Fig. 2.

This paper provides a comprehensive survey on RGB-D based SOD, which aims to thoroughly cover various aspects of the models for this task and provide insightful discussions on the challenges and open directions for future work. We also review another related topic, i.e., light field SOD, in which the light field can provide more information (including focal stack, all-focus images and depth maps) to boost the performance of salient object detection. Further, we provide a comprehensive comparison to evaluate existing RGB-D based SOD models and provide their main advantages.

1.1 Related Reviews and Surveys

There are several surveys that are closely related to salient detection. For example, Borji et al. [2] provided a quantitative evaluation of 35 state-of-the-art non-deep saliency detection methods. Cong et al. [17] reviewed several different saliency detection models, including RGB-D based SOD, co-saliency detection, and video SOD. Zhang et al. [155] provided an overview of co-saliency detection, and review the history of co-saliency detection and summarize several benchmark algorithms in this field. Han et al. [47] reviewed the recent progress in SOD including models, benchmark datasets, and evaluation metrics, as well as discussed the underlying connection among object detection, SOD, and category-specific object detection. Nguyen et al. [94] reviewed various works related to saliency applications and provided insightful discussions on role of saliency in these. Borji et al. [1] provided a comprehensive review of recent progress in SOD and discuss some related works including generic
scene segmentation, saliency for fixation prediction, and object proposal generation. Fan et al. [28] provided a comprehensive evaluation of several state-of-the-art CNNs-based SOD models, and proposed a high quality SOD dataset, termed SOC (Details can be found at: http://dpfan.net/socbenchmark/). Zhao et al. [178] reviewed various deep learning-based object detection models and algorithms in detail, and also reviewed various specific tasks including SOD works. Wang et al. [138] focused on reviewing deep learning based SOD models. Different from previous SOD surveys, in this paper, we focus on reviewing the existing RGB-D based SOD models and benchmark RGB-D datasets.

1.2 Contributions
The main contributions are summarized as follows:
- We provide the first systematic review of RGB-D based SOD models from different perspectives. We summarize existing RGB-D SOD models into traditional or deep methods, fusion-wise methods, single-stream/multi-stream methods, and attention-aware methods.
- We review nine RGB-D benchmark datasets that are popularly used in this field, and we provide details for each dataset. Moreover, we provide a comprehensive as well as an attribute-based evaluation of several representative RGB-D based SOD models.
- We supply the first collection and review of the related light field SOD models and benchmark datasets.
- We thoroughly investigate several challenges for RGB-D based SOD field and the relation of SOD with other topics, shedding light on potential directions for future research.

1.3 Organization
In Sec. 2, we review existing RGB-D based models from different aspects. In Sec. 3, we summarize and provide details for current benchmark datasets for RGB-D salient object detection. In Sec. 4, we conduct a comprehensive review of light field SOD models and benchmark datasets. In Sec. 5, we provide a comprehensive and attribute-based evaluation of several representative RGB-D based models. We then discuss challenges and open directions of this field in Sec. 6. Finally, we conclude this paper in Sec. 7.

2 RGB-D based SOD Models
Over the past few years, several RGB-D based SOD methods have been developed and obtained promising performance. These models are summarized in Tables 1, 2, 3 and 4. The complete benchmark can be found at http://dpfan.net/d3netbenchmark/.

To review these RGB-D based SOD models in detail, we introduce them from the following perspectives: (1) Traditional/deep models; (2) fusion-wise models, (3) single-stream/multi-stream models, and (4) attention-aware models.

2.1 Traditional/Deep Models

Traditional Models. With depth cues, we can often explore several useful attributes, such as boundary cues, shape attributes, surface normals, etc., to boost the identification of salient objects in complex scenes. Over the past several years, many traditional RGB-D models based on handcrafted features have been developed [7, 14, 16, 22, 24, 27, 38, 41, 44, 45, 63, 79, 99, 111, 114, 121, 190]. For example, the early work [16] focused on modeling the interaction between layout and shape features generated from the RGB image and depth map. Besides, the representative work [99] developed a novel multi-stage RGB-D model, and constructed the first large-scale RGB-D benchmark dataset, termed NLPR.

Deep Models. However, the above-mentioned methods suffer from unsatisfactory SOD performance due to the limited expression ability of handcrafted features. To address this, several studies have turned to deep neural networks (DNNs) to fuse RGB-D data [3, 6, 8, 9, 11, 18, 42, 52, 54, 67, 71, 78, 83, 85, 101, 103, 106, 109, 114, 132, 160, 167, 173, 186, 188]. These models can learn high-level representations to explore complex correlations across RGB images and depth cues for improving SOD performance. We review some representative works in details as follows.
- DF [109] develops a novel convolutional neural network (CNN) to integrate different low-level saliency cues into hierarchical features, for effectively locating salient regions in RGB-D images. This was the first CNN-based model for the RGB-D SOD task. However, it utilizes a shallow architecture to learn the saliency map.
- PCF [4] presents a complementarity-aware fusion module to integrate cross-modal and cross-level feature representations, which can effectively exploit complementary information using cross-modal/level connections and modal/level-wise supervisions explicitly to decrease fusion ambiguity.
- CTMF [46] develops a computational model to identify salient objects from RGB-D scene, utilizing CNNs to learn high-level representations for RGB
Tab. 1  Summary of RGB-D based SOD methods (published from 2012 to 2016).

| #  | Year | Method | Pub. | Dataset | Description |
|----|------|--------|------|---------|-------------|
| 1  | 2012 | DM     | ECCV | NUS-3D Salient dataset | Models the correlation between saliency and depth by approximating the joint density using Gaussian Mixture Models. |
| 2  | 2012 | RCM    | ICCCSE | 300 objects | Develops a region contrast based SOD model with depth cue. |
| 3  | 2013 | LS     | BMVC | GIT | Extends the dissimilarity framework to model the joint interaction between depth cues and RGB images. |
| 4  | 2013 | RC     | BMVC | UW dataset, Berkeley 3D dataset | Derives RGB-D saliency by formulating a 3D saliency model based on the region contrast of the scene and fuses it using SVM. |
| 5  | 2013 | SOS    | NEURO | 300 RGB-D images | Incorporates depth cues for salient object segmentation by suppressing background regions. |
| 6  | 2014 | SRDS   | ICDSP | STERE | Integrates depth and depth weighted color contrast with spatial compactness of color distribution. |
| 7  | 2014 | LHM    | ECCV | NLRP | Uses a multi-stage RGB-D algorithm to combine both depth and appearance cues to segment salient objects. |
| 8  | 2014 | DESM   | ICIMCS | DES | Combines three saliency cues: color contrast, spatial bias, and depth contrast. |
| 9  | 2014 | ACSD   | ICIP | 1382 images | Measures a point’s saliency by how much it stands out from the surroundings, and has two priors (regions nearer to viewers are more salient and salient objects tend to be located at the center). |
| 10 | 2015 | GP     | CVPRW | NLRP, NJUD | Explores orientation and background priors for detecting salient objects, and uses PageRank and MRPs to optimize the saliency maps. |
| 11 | 2015 | SFP    | ICMICS | NLRP, NJUD | Develops a RGB-D based SOD approach using saliency fusion and propagation. |
| 12 | 2015 | DIC    | TVC | 103 stereo pairs | Fuses the saliency maps from color and depth to generate a noise-free salient patch, and utilizes random walk algorithm to infer the object boundary. |
| 13 | 2015 | SRD    | ICRA | GT, MER | Designs a graph-based segmentation to identify homogeneous regions using color and depth cues. |
| 14 | 2015 | MGMR   | ICIIP | NLRP | Designs a mutual guided manifold ranking strategy to achieve SOD. |
| 15 | 2015 | SF     | CAC | NLRP | Proposes to automatically select discriminative features using decision trees for better performance. |
| 16 | 2016 | PRC    | ACCESS | NLRP, NJUD | Saliency fusion and progressive region classification are used to optimize depth-aware saliency models. |
| 17 | 2016 | LBR    | CVPR | NLRP, NJUD | Uses a local background enclosure to capture the spread of angular directions. |
| 18 | 2016 | SE     | ICME | NLRP, NJUD | Utilizes cellular automatons to propagate the initial saliency map and then generate the final saliency prediction result. |
| 19 | 2016 | DCMC   | SPL | NJUD | Develops a new measure to evaluate the reliability of depth maps for reducing the influence of poor-quality of depth maps on saliency detection. |
| 20 | 2016 | BF     | ICPR | IIC-cyN/IVC 3D Gaze | Fuses contrast features from RGB and depth images with a Bayesian framework. |
| 21 | 2016 | DCI    | ICASSP | STERE, NJUD | Adopts the original depth map to subtract the fitted surface for generating a contrast increased map. |
| 22 | 2016 | DSF    | ICASSP | NLRP, NJUD | Develops a multi-stage depth-aware saliency model for SOD. |
| 23 | 2016 | GM     | ACCV | NLRP | Combines color and depth-based contrast features using a generative mixture model. |

Images and depth cues, and exploit the complementary correlations and joint representation simultaneously. Besides, this model transfers the structure of the model from the source domain (i.e., RGB images) to be applicable to the target domain (i.e., depth maps).

- **CPFP [173]** proposes a contrast-enhanced network to produce the enhanced map, and presents a fluid pyramid integration module to effectively fuse cross-modal information via a hierarchical manner. Besides, considering the fact that depth cues easily suffer from noisy, a feature-enhanced module is proposed to learn an enhanced depth cue for boosting the SOD performance. It is worth noting that this is an effective solution.

- **UC-Net [160]** proposes a probabilistic RGB-D based SOD network via conditional variational autoencoders (VAE) to model human annotation uncertainty. It generates multiple saliency maps for each input image by sampling in the learned latent space. This was the first work to investigate uncertainty in RGB-D based SOD, and was inspired by the data labeling process. This method leverages the diverse saliency maps to improve the final SOD performance.

2.2 Fusion-wise Models

For RGB-D based SOD models, it is important to effectively fuse RGB images and depth maps. The existing fusion strategies can be grouped into three categories, including 1) early fusion, 2) multi-scale fusion, and 3) late fusion. We provide details for each fusion strategy as follows.

**Early Fusion.** Early fusion-based methods can follow one of two veins: 1) RGB images and depth maps are directly integrated to form a four-channel input [83, 99, 111, 117, 117]. This is denoted as “input fusion” (shown in Fig. 3); 2) RGB and depth images are first fed into each independent network and their low-level representations are combined as joint representations, which are then fed into a subsequent network for further saliency map prediction [109]. This is denoted as “early feature fusion” (shown in Fig. 3).

**Late Fusion.** Late fusion-based methods can also
be further divided into two families: 1) Two parallel network streams are adopted to learn high-level features for RGB and depth data, respectively, which are concatenated and then used for generating the final saliency prediction [24, 46, 132]. This is denoted as “later feature fusion” (shown in Fig. 3); 2) Two parallel network streams are used to obtain the independent saliency map for RGB images and depth cues, and then the two saliency maps are concatenated to obtain a final prediction map [25]. This is denoted as “late result fusion” (shown in Fig. 3).

Multi-scale Fusion. To effectively explore the correlations between RGB images and depth maps, several methods propose a multi-scale fusion strategy [9, 13, 36, 42, 71, 72, 98, 171]. The first category learn the cross-modal interactions and then fuse them into a feature learning network. For example, Chen et al. [9] developed a multi-scale multi-path fusion network to integrate RGB images and depth maps, with a cross-modal interactions (termed as MMCI) module. This method introduces cross-modal interactions into multiple layers, which can empower additional gradients for enhancing the learning of the depth stream as well as explore complementarity across low-level and high-level representations. The second category fuse the features from RGB images and depth maps in different layers and then integrate them into a decoder network (e.g., skip connection) to produce the final saliency detection map (as shown in Fig. 3). Some representative works are briefly discussed as follows.

- **ICNet** [71] proposes an information conversion module to convert high-level features via an interactive manner. In this model, a cross-modal depth-weighted combination (CDC) block is proposed to enhance RGB features with depth features at different levels.
- **DPANet** [13] uses a gated multi-modality attention (GMA) module to exploit long-range dependencies. The GMA module can extract the most discriminative features by utilizing a spatial attention mechanism. Besides, this model controls the fusion rate of the cross-modal information using a gate function, which can reduce some effects brought by the unreliable depth cues.
- **BiANet** [171] develops a multi-scale bilateral attention module (MBAM) to capture better global information in multiple layers.
Fig. 3 Comparison of three fusion strategies that explore the correlation between RGB images and depth maps for RGB-D based SOD. These include: 1) Early fusion; 2) Late fusion; 3) Multi-scale fusion.

Tab. 3 Summary of RGB-D based SOD models published in 2019

| No. | Year | Method | Pub. | Dataset | Description |
|-----|------|--------|------|---------|-------------|
| 42  | 2019 | SSRC  [53] | NEURO | NLPR, NJUD, STERE, LFSD | Uses a single stream recurrent convolution neural network with a four-channel input and DRCNN subnetwork |
| 43  | 2019 | MLP  [33] | SPL  | NJUD | Designs a salient object-aware data augmentation method to expand the training set |
| 44  | 2019 | TSRN  [80] | ICIP | NJUD | Designs a fusion refinement module to integrate output features from different modalities and resolutions |
| 45  | 2019 | DIL  [26] | MTAP | NLPR, NJUD | Designs a consistency integration strategy to generate an image pre-segmentation result that is consistent with the depth distribution |
| 46  | 2019 | CAFM  [185] | TIP | NUS [65], NCTU [69] | Utilizes a content-aware fusion module to integrate global and local information |
| 47  | 2019 | PDNet  [188] | TIP | NLPR, NJUD | Adopt a prior-model guided master network to process RGB information, which is pre-trained on the conventional RGB dataset to overcome the limited size |
| 48  | 2019 | MMCI  [5] | PR  | NLPR, NJUD, STERE | Improves the traditional two-stream architecture by diversifying the multi-modal fusion paths and introducing cross-modal interactions in multiple layers |
| 49  | 2019 | DCA  [103] | TIP | LFSD | Enforces spatial consistency by constructing an optimization model, and the saliency value of each superpixel is updated by exploiting the intrinsic relevance of similar regions |
| 50  | 2019 | TANet  [6] | TIP | NLPR, NJUD, STERE | Uses a three-stream multi-modal fusion framework to explore cross-modal modality complementarity in both the bottom-up and top-down processes |
| 51  | 2019 | DCNRF  [8] | TCYB | NLPR, NJUD | Formulates a CNN-based cross-modal transfer learning problem for depth-induced SOD, and uses a dense cross-level feedback strategy to exploit cross-level interactions |
| 52  | 2019 | DGT  [18] | TCYB | NLPR, NJUD, STERE | Utilizes a content-aware fusion module to integrate global and local information |
| 53  | 2019 | LSP  [5] | arXiv | NLPR, NJUD, STERE | Designs an RGB-D system with three key components, including modal-specific representations learning, complementary information selection, and cross-modal complementation fusion |
| 54  | 2019 | AFNet  [132] | ACCESS | NLPR, NJUD, STERE, LFSD, DES | Learns a switch map that is used to adaptively fuse the predicted saliency maps from the RGB and depth modality |
| 55  | 2019 | EPM  [62] | ACCESS | NLPR, NJUD, STERE, LFSD, DES | Develops an effective propagation mechanism for RGB-D co-saliency detection |
| 56  | 2019 | CPFP  [173] | CVPR | NLPR, NJUD, STERE | Uses a contrast-enhanced network to obtain the one-channel enhanced map, and designs a fluid pyramid integration module to fuse cross-modal cross-level features in a pyramid style |
| 57  | 2019 | DMRA  [101] | ICCV | NJUD, NLPR, STERE, LFSD, DUT-RGBD, DES, SSD | Designs a depth-induced multiscale recurrent attention network for SOD, including a depth refinement block and a recurrent attention module |
| 58  | 2019 | DSD  [25] | JVCIR | NJUD, NLPR, STERE, DES, SSD | Uses a saliency fusion network to adaptively fuse both the color and depth saliency maps |

- **JL-DCF** [42] treats a depth image as a special case of a color image and employs a shared CNN for both RGB and depth feature extraction. It also proposes a densely-cooperative fusion strategy to effectively combine the learned features from different modalities.
- **BBS-Net** [36] uses a bifurcated backbone strategy (BBS) to split the multi-level feature representations into teacher and student features, and develops a depth-enhanced module (DEM) to explore informative parts in depth maps from the spatial and channel views.

### 2.3 Single-stream/Multi-stream Models

**Single-stream Models.** Several RGB-D based SOD works [52, 83, 109, 114, 117, 173, 188] focus on a single-stream architecture to achieve saliency prediction. These models often fuse RGB images and depth information in the input channel or feature learning part. For example, MDSF [117] develops a multi-scale discriminative saliency fusion framework as the SOD model, in which four types of features in three levels are computed and then fused to obtain the final
| No. | Year | Method | Pub. | Dataset | Description |
|-----|------|--------|------|---------|-------------|
| 59  | 2020 | DPANet [13] | arXiv | NJUD, NLPR, SISTERE, LFSD, DES, DUT-RGBD, SIP | Uses a saliency-oriented depth perception module to evaluate the potentiality of depth maps and reduce effects of contamination |
| 60  | 2020 | CAP [72] | arXiv | NJUD, NLPR, SISTERE, SSD, DES | Utilizes depth cues as training prior to facilitate SOD |
| 61  | 2020 | SSDP [14] | arXiv | STERE, LFSD, NJUD, NLPR, SIP, DUT-RGBD, DES | Makes use of existing labeled RGB saliency datasets together with unlabeled RGB-D data to boost SOD performance |
| 62  | 2020 | AttNet [186] | IVC | NJUD, NLPR, SISTERE, LFSD, DES | Deploys attention maps to boost the salient objects’ location and pays more concern to the appearance information |
| 63  | 2020 | GFNet [86] | Neuro | NLPR, NJU, SISTERE, DES, SIP | Uses an adaptive gated fusion module via a GAN to obtain a better fused saliency map from RGB images and depth cues |
| 64  | 2020 | CoCNN [78] | Pr | SISTERE, NJUD | Fuses color and disparity features from low to high layers in a unified deep model |
| 65  | 2020 | cmSalGAN [59] | TMM | NJUD, NLPR, STERE | Aims to learn an optimal view-invariant and consistent pixel-level representation for both RGB and depth images using an adversarial learning framework |
| 66  | 2020 | PGGF [16] | ACCESS | NJUD, NLPR, LFSD, STERE, DES | Leverages powerful representations learned from large-scale RGB datasets to boost the model ability |
| 67  | 2020 | BiANet [171] | TIP | NJUD, NLPR, SISTERE, SSD, DES, SIP | Uses a bilateral attention module (BAM) to explore rich foreground and background information from depth maps |
| 68  | 2020 | ASFNet [57] | TCYB | NJUD, NLPR, SITERE, LFSD | Integrates the attention-sustained complements from RGB-D images and introduces a global semantic constraint using adversarial learning |
| 69  | 2020 | Triple-Net [34] | SPL | NJUD, STERE, LFSD, DES | Uses a triple-complementary network for RGB-D based SOD |
| 70  | 2020 | ICNet [71] | TIP | NJUD, NLPR, SITERE, LFSD, DES | Uses a novel information conversion module to fuse high-level RGB and depth features in an interactive and adaptive way |
| 71  | 2020 | SDF [3] | TIP | NLPR, NJUD, DDC, LFSD | Proposes an exemplar-driven method to estimate relatively-trustworthy depth maps, and uses a selective deep saliency fusion network to effectively integrate RGB images, original depths, and newly estimated depths |
| 72  | 2020 | GFNet [186] | SPL | NJUD, NLPR | Designs a gate fusion block to regularize feature fusion |
| 73  | 2020 | RGBS [84] | MTAP | NJUD, NLPR, SIP, STERE | Utilizes a GAN to generate the saliency map |
| 74  | 2020 | D’Net [55] | TNNLS | NJUD, NLPR, SIB, LFSD, DES, GIT, STERE, SIP | Uses a depth degrader unit (DDU) and a three-stream feature learning module to employ low-quality depth cue filtering and cross-modal feature learning, respectively |
| 75  | 2020 | JL-DCF [12] | CVPR | NJUD, NLPR, SITERE, DES, LFSD, SIP | Uses a joint learning strategy and a density-cooperative fusion module to achieve better SOD performance |
| 76  | 2020 | A2dele [106] | CVPR | NJUD, NLPR, DUT-RGBD, DES, STERE | Employs a depth distiller to explore ways of using network prediction and attention as two bridges to transfer deep knowledge to RGB images |
| 77  | 2020 | SSF [16] | CVPR | NJUD, NLPR, DUT-RGBD, DES, STERE | Designs a complimentary interaction module to select useful representations from the RGB and depth images and then integrate cross-modal features |
| 78  | 2020 | S’tMa [82] | CVPR | NJUD, NLPR, SITERE, LFSD, DUT-RGBD, DES, SSD | Fuses multi-modal information via self-attention and each other’s attention strategies, and reweights the mutual attention term to filter out unreliable information |
| 79  | 2020 | UC-Net [190] | CVPR | NJUD, NLPR, SIB, LFSD, DES | Uses a probabilistic RGB-D saliency detection network via a conditional VAE to generate multiple saliency maps |
| 80  | 2020 | CMWNNet [72] | ECCV | NJUD, NLPR, SITERE, DES, LFSD, SSD, SIP | Exploits feature interactions using three cross-modal fusion-weighting modules to improve SOD performance |
| 81  | 2020 | HDFNet [58] | ECCV | NJUD, NLPR, SITERE, DES, LFSD, SSD, SIP, DUT-RGBD | Designs a hierarchical dynamic filtering network to effectively make use of cross-modal fusion information |
| 82  | 2020 | CAS-GNN [38] | ECCV | NJUD, NLPR, SITERE, LFSD, SES, SSD | Designs cascaded graph neural networks to exploit useful knowledge from RGB and depth images for building powerful feature embeddings |
| 83  | 2020 | CMMS [58] | ECCV | NJUD, NLPR, SITERE, LFSD, SSD, DUT-RGBD | Proposes a cross-modality feature modulation module to enhance feature representations and an adaptive feature selection module to gradually select saliency-related features |
| 84  | 2020 | DANet [177] | ECCV | NJUD, NLPR, DUT-RGBD, DES, SSD, SIP | Develops a single-stream network combined with a depth-enhanced dual attention to achieve real-time SOD |
| 85  | 2020 | CoNet [35] | ECCV | NJUD, NLPR, DUT-RGBD, DES, SSD, LFSD, SIP | Develops a collaborative learning framework for RGB-D based SOD, and three collaborators (edge detection, coarse salient object detection and depth estimation) are utilized to jointly boost the performance |
| 86  | 2020 | BBS-Net [36] | ECCV | NJUD, NLPR, SITERE, LFSD, DES, SSD, SIP | Uses a bifurcated backbone strategy to learn teacher and student features, and utilizes a depth-enhanced module to excavate informative parts of depth cues |

**Saliency Maps and Depth Images (Saliency Map, Depth Map)**: The saliency map and depth images play a crucial role in RGB-D salient object detection. A saliency map provides the likelihood of an object being salient, while a depth image contains depth information. Utilizing both modalities can significantly improve the accuracy of salient object detection. This dual-modality approach allows for better understanding of the scene depth and object location.

**Multi-stream Models**: Two-stream models [101, 132, 186] consist of two independent branches that process RGB images and depth cues, respectively, and often generate different high-level features or saliency maps and then incorporate them in the middle stage or end of the two streams. It is worth noting that most recent deep learning-based models [4, 5, 8, 9, 13, 59, 67, 71, 85, 106] utilize this two-stream architecture with several models capturing the correlations between RGB and depth images to achieve superior performance.
RGB images and depth cues across multiple layers. Moreover, some models utilize a multi-stream structure [6, 34] and then design different fusion modules to effectively fuse RGB and depth information to exploit their correlations.

### 2.4 Attention-aware Models

Existing RGB-D based SOD methods often treat all regions using the extracted features equally, while ignoring the fact that different regions can have different contributions to the final prediction map. These methods are easily affected by cluttered backgrounds. In addition, some methods either regard the RGB images and depth maps as having the same status or overly rely on depth information. This prevents them from considering the importance of different domains (RGB images or depth cues). To overcome this, several methods introduce attention mechanisms to weight the importance of different regions or domains.

- **ASIF-Net** [67] captures complementary information from RGB images and depth cues using an interweaved fusion, and weights the saliency regions though a deeply supervised attention mechanism.
- **AttNet** [186] introduces attention maps for differentiating of salient objects and background regions to reduce negative influence of some low-quality depth cues.
- **TANet** [6] formulates a multi-modal fusion framework using RGB images and depth maps from the bottom-up and top-down views, and proposes a channel-wise attention module to effectively fuse the complementary information from different modalities and levels.
- **ACCF** [11] proposes a top-down RGB-D fusion network, in which an attention-induced cross-modal cross-level fusion module is adopted to extract informative features from each modality at different levels.
2.5 Open-source Implementations

We summarize the open-source implementations of RGB-D based SOD models reviewed in the survey. The implementations and hyperlinks of the source codes of these models are provided in Tab 5.

3 RGB-D Datasets

With the rapid development of RGB-D based SOD, various datasets have been constructed over the past several years. Tab 6 summarizes nine popular RGB-D datasets, and Fig. 4 shows examples of images (including RGB images, depth maps, and annotations) for these datasets. Moreover, we provide the details for each dataset as follows.

- **STERE** [96]. The authors first collected 1250 stereoscopic images from Flickr \(^1\), NVIDIA 3D Vision Live \(^2\), and Stereoscopic Image Gallery \(^3\). The most salient objects in each image were annotated by three users. All annotated images were then sorted based on the overlapping salient regions and the top 1000 images were selected to construct the final dataset. This is the first collection of stereoscopic images in this field.

- **GIT** [16] consists of 80 color and depth images, which were collected using a mobile-manipulator robot in a real-world home environment. Moreover, each image is annotated based on the pixel-level segmentation of the objects.

- **DES** [14] consists of 135 RGB-D indoor images, which were taken by Kinect with a resolution of 640 × 640. When collecting this dataset, three users were asked to label the salient object in each image, and then the overlapping areas of the labeled object were regarded as the ground truth.

- **NLPR** [99] consists of 1000 RGB images and their corresponding depth maps, which were obtained by a standard Microsoft Kinect. This dataset includes a series of outdoor and indoor locations, e.g., offices,

---

---

1http://www.flickr.com/
2http://photos.3dvisionlive.com/
3http://www.stereophotography.com/
supermarkets, campuses, streets, and so on.

- LFSD [75] includes 100 light fields collected using a Lytro light field camera, and consists of 60 indoor and 40 outdoor scenes. To label this dataset, three individuals were asked to manually segment salient regions, and then the segmented results were deemed ground truth when the overlap of the three results was over 90%.

- NJUD [63] consists of 1985 stereo image pairs, and these images were collected from the internet, 3D movies, and photographs that are taken by a Fuji W3 stereo camera.

- SSD [189] was constructed using three stereo movies and includes indoor and outdoor scenes. This dataset includes 80 samples, and each image has the size of $960 \times 1080$.

- DUT-RGBD [103] consists of 800 indoor and 400 outdoor scenes with their corresponding depth images. This dataset includes several challenging factors, i.e., multiple or transparent objects, complex backgrounds, similar foregrounds and backgrounds, and low-intensity environments.

- SIP [34] consists of 929 annotated high-resolution images, with multiple salient persons in each image. In this dataset, depth maps were captured using a real smartphone (i.e., Huawei Mate10). Besides, it is worth noting that this dataset covers diverse scenes, and various challenging factors, and is annotated with pixel-level ground truths.

### 4 Saliency Detection on Light Field

#### 4.1 Light Field SOD Models

Existing works for SOD can be grouped into three categories according to the input data type, including RGB SOD, RGB-D SOD, and light field SOD [165]. We have already reviewed RGB-D based SOD models, in which depth maps provide layout information to improve SOD performance to some extent. However, inaccurate or low-quality depth maps often decrease the performance. To overcome this issue, light field SOD methods have been proposed to make use of rich information captured by the light field. Specifically, light field data contains an all-focus image, a focal stack, and a rough depth map [103]. A summary of related light field SOD works is provided in Tab. 7. Further, to provide an in-depth understanding of these models, we also review them in more detail as follows.

#### Traditional/Deep Models

The classic models for light field SOD often use superpixel-level hand-crafted features [74–76, 103, 113, 128, 133, 143, 163, 164]. Early work [75, 76] showed that the unique refocusing capability of light fields can provide useful focusness, depth, and objectness cues, and further proposed several SOD models using light field data. For example, Zhang et al. [163] utilized a set of focal slices to compute the background prior, and then incorporate it with the location prior for SOD. Wang et al. [128] proposed a two-stage Bayesian fusion model to integrate multiple contrasts for boosting SOD performance. Recently, several deep learning-based light field SOD models [104, 105, 137, 162, 165, 166] have been developed and obtained remarkable performance. Besides, in [137], an attentive recurrent CNN was developed to fuse all focal slices, while the data diversity was increased using adversarial examples to enhance model robustness. Zhang et al. [166] developed a memory-oriented decoder for light field SOD, which fuses multi-level features in a top-down strategy using high-level information to guide low-level feature selection. LFNet [165] employs a new integration module to fuse features from light field data according to their contributions and captures the spatial structure of a scene to improve SOD performance.

#### Refinement based Models

Several refinement strategies have been used to enforce neighboring constraints or reduce the homogeneity of multiple modalities for SOD. For example, in [74], the saliency

---

**Tab. 6** Statistics of nine RGB-D benchmark datasets in terms of year (Year), publication (Pub.), dataset size (Size), number of objects in the images (#Obj.), type of scene (Types), depth sensor (Sensor), and resolution (Resolution). See Sec. 3 for more details on each dataset. These datasets can be downloaded from our website: http://dpfan.net/d3netbenchmark/.

| # | Dataset | Year | Pub. | Size | #Obj. | Types | Sensor | Resolution |
|---|---|---|---|---|---|---|---|---|
| 1 | STERE [96] | 2012 | CVPR | 1000 | One | Home environment | Stereo camera+shift flow | $[251 \sim 1200] \times [222 \sim 900]$ |
| 2 | GIT [10] | 2013 | BMVC | 80 | Multiple | Indoor | Microsoft Kinect | $480 \times 480$ |
| 3 | DES [14] | 2014 | ICCV | 135 | One | Indoor | Microsoft Kinect | $480 \times 480$ |
| 4 | NLPR [69] | 2014 | ECCV | 1000 | Multiple | Indoor/outdoor | Microsoft Kinect | $480 \times 480, 480 \times 640$ |
| 5 | LFSD [75] | 2014 | CVPR | 1985 | One | Indoor/outdoor | Lytro Illum camera | $360 \times 360$ |
| 6 | NJUD [63] | 2014 | ICIP | 1200 | Movies/internet/photo | Fuji W3 camera+optical flow | Sun’s optical flow | $[231 \sim 1213] \times [274 \sim 828]$ |
| 7 | SSD [189] | 2017 | ICCV | 828 | Multiple | Movies | Huawei Mate10 | $960 \times 1080$ |
| 8 | DUT-RGBD [103] | 2019 | ICCV | 1200 | Multiple | Indoor/outdoor | Microsoft | $400 \times 600$ |
| 9 | SIP [34] | 2020 | TNNLS | 929 | Multiple | Person in the wild | Huawei Mate10 | $992 \times 744$ |
Tab. 7 Summary of popular light field SOD methods.

| No | Year | Method | Pub. | Dataset | Description |
|----|------|--------|------|---------|-------------|
| 1  | 2014 | LFS [75] | CVPR | LFSD    | Develops the first light-field saliency detection algorithm to employ the objectness and focusness cues based on the refocusing capability of the light field |
| 2  | 2015 | WSC [74] | CVPR | LFSD    | Uses a weighted sparse coding framework to learn a saliency/non-saliency dictionary |
| 3  | 2015 | DILF [163] | IJCAI | LFSD    | Incorporates depth contrast to complement the disadvantage of color and conducts focusness-based background priors to boost the saliency detection performance |
| 4  | 2016 | RL [113] | ICASSP | LFSD    | Utilizes the inherent structure information in light field images to improve saliency detection |
| 5  | 2017 | MA [164] | TOMM | LFSD    | Integrates multiple saliency cues extracted from light field images using a random-serach-based weighting manner |
| 6  | 2017 | BIF [128] | NPL | LFSD    | Integrates color-based contrast, depth-induced contrast, focusness map of foreground slice, and background weighted depth contrast are fused using a two-stage Bayesian integration framework |
| 7  | 2017 | LFS [76] | TPAMI | LFSD    | An extension of [75] |
| 8  | 2017 | RLM [70] | ICIVC | LFSD    | Utilizes the light field relative location measurement for SOD on light field images |
| 9  | 2018 | SGDC [133] | CVPR | LFSD    | Designs a salience-guided depth optimization framework for multi-layer light field displays |
| 10 | 2018 | DCA [102] | PiO | LFSD    | Proposes a graph model depth-induced cellular automata to optimize saliency maps using light field data |
| 11 | 2019 | DLLF [135] | ICCV | DUTLF-FS, LFSD | Utilizes a recurrent attention network to fuse each slice from the focal stack to learn the most informative features |
| 12 | 2019 | DLSD [104] | IJCAI | DUTLF-MV | Formulates saliency detection into two subproblems, including 1) light field synthesis from a single view and 2) light-field-driven saliency detection |
| 13 | 2019 | MoI [196] | NIPS | UTLF-FS | Uses a memory-oriented decoder for light field SOD |
| 14 | 2020 | ERNet [105] | AAAI | DUTLF-FS, HFUT, LFSD | Uses an asymmetrical two-stream architecture to overcome computation-intensive and memory-intensive challenges in a high dimension of light field data |
| 15 | 2020 | DCA [103] | TIP | LFSD    | Uses an asymmetrical two-stream architecture to overcome computation-intensive and memory-intensive challenges in high dimension of light field data |
| 16 | 2020 | RDFD [143] | MTAP | LFSD    | Defines a region-based depth feature descriptor extracted from the light field focal stack to facilitate low- and high-level cues for saliency detection |
| 17 | 2020 | LFNet [165] | TIP | DUTLF-FS, LFSD, HFUT | Utilizes a light field refinement module and a light field integration module to effectively integrate multiple cues (i.e., focusness, depths and objectness) from light field images |
| 18 | 2020 | LFDCN [182] | TIP | Lytro Illum, LFSD, HFUT | Uses a deep convolutional network based on the modified DeepLab-v2 model to explore spatial and multi-view properties of light field images for saliency detection |

There are four representative datasets widely-used in existing light field SOD models. We describe the details for each dataset as follows.

- **LFSD** [75] consists of 100 light fields of different scenes with 360 × 360 spatial resolution, captured using a Lytro light field camera. This dataset contains 60 indoor and 40 outdoor scenes, and most scenes consist of only one salient object. Besides, three individuals were asked to manually segment salient regions in each image, and then the ground truth was determined when all three segmentation results had an overlap of over 90%.

- **HFUT** [164] consists of 255 light fields captured using a Lytro camera. In this dataset, most scenes contain multiple objects that appear within different locations and scales under complex background clutter.

- **DUTLF-FS** [137] consists of 1465 samples, 1000 of which are as training set while the remaining 465 images are as testing set. The resolution of each image is 600 × 400. This dataset contains several challenges, e.g., lower contrast between salient objects and cluttered background, multiple disconnected salient objects, and dark or strong light conditions.

- **DUTLF-MV** [104] consists of 1580 samples, in which 1100 samples are for training and the remaining is for testing. Images were captured by a Lytro Illum camera, and each light field consists of multi-view images and a corresponding ground truth.

- **Lytro Illum** [162] consists of 640 light fields and the corresponding per-pixel ground-truth saliency maps. It includes several challenging factors, e.g., inconsistent illumination conditions, and small salient

---

1. https://sites.duke.edu/nianyi/publication/saliency-detection-on-light-field/
2. https://github.com/pencilzhang/HFUT-Lytro-dataset
3. https://github.com/OIPLab-DUT/ICCV2019_DeepLightField_Saliency
4. https://github.com/OIPLab-DUT/IJCAI2019_Deep-Light-Field-Driven-Saliency-Detection-from-A-Single-View
5. https://github.com/pencilzhang/MAC-light-field-saliency-net
objects existing in a similar or cluttered background.

5 Model Evaluation and Analysis

5.1 Evaluation Metrics

We briefly review several popular metrics for SOD evaluation as follows.

• **MAE.** This is the *mean absolute error* (MAE) \[100\] between a prediction saliency map \(S\) and a ground truth \(G\) for all pixels, which can be defined by

\[
MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |S_{i,j} - G_{i,j}|, \tag{1}
\]

where \(W\) and \(H\) denote the width and height of the map, respectively. MAE values are normalized to \([0,1]\).

• **S-measure** \((S_\alpha)\). To capture the importance of image structural information, \(S_\alpha\) \[29\] is used to assess the structural similarity between the regional perception \((S_r)\) and object perception \((S_o)\). Thus, \(S_\alpha\) can be defined by

\[
S_\alpha = \alpha \ast S_o + (1 - \alpha) \ast S_r, \tag{2}
\]

where \(\alpha \in [0,1]\) is a trade-off parameter. Here, we set \(\alpha = 0.5\) as the default setting as suggested by Fan *et al.* \[29\].

• **E-measure** \((E_\phi)\). \(E_\phi\) \[30\] was proposed based on cognitive vision studies to capture image-level statistics and their local pixel matching information. Thus, \(E_\phi\) can be defined by

\[
E_\phi = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \phi_{FM}(i,j), \tag{3}
\]

where \(\phi_{FM}\) denotes the enhanced-alignment matrix \[30\].

• **F-measure** \((F_\beta)\). \(F_\beta\) is popular metric and has been widely applied to evaluate the performance of SOD. Inspired by \[2\] and \[34\], we use fixed \([0,255]\) thresholds to compute this metric. Finally, \(F_\beta\) is calculated by

\[
F_\beta = \frac{1 + \beta^2}{\beta^2 \ast P + R}, \tag{4}
\]

where \(\beta^2\) is set to 0.3 to emphasize the precision.

• **PR Curve.** As proposed in \[2\], a saliency map \(S\) is divided using different thresholds \((i.e., it changes from 0 to 255)\). For each threshold, we first calculate a pair of recall and precision scores, and then combine them to obtain a precision-recall curve that describes the performance of the model at the different thresholds.

5.2 Performance Comparison and Analysis

5.2.1 Overall Evaluation

To quantify the performance of different models, we conduct a comprehensive evaluation of 24 representative RGB-D based SOD models, including 1) 9 traditional methods: LHM \[99\], ACSD \[63\], DESM \[14\], GP \[111\], LBE \[41\], DCMC \[22\], SE \[44\], CDCP \[190\], CDB \[79\]; and 2) 15 deep learning-based methods: DF \[109\], PCF \[4\], CTMF \[46\], CPFP \[173\], TANet \[6\], AFNet \[132\], MMCI \[9\], DMRA \[101\], D3Net \[34\], SSF \[167\], A2dele \[106\], S2MA \[82\], ICNet \[71\], JL-DCF \[42\], and UC-Net \[160\]. We report the mean values of \(S_\alpha\) and MAE across the five datasets \((i.e., \text{STERE} [96], \text{NLPR} [99], \text{LFSD} [75], \text{DES} [14], \text{and SIP} [34])\). We report the mean values of \(S_\alpha\) and MAE across the five datasets \((i.e., \text{STERE} [96], \text{NLPR} [99], \text{LFSD} [75], \text{DES} [14], \text{and SIP} [34])\) in each model. Note that these better models are shown in the upper left corner \((i.e., with a larger \(S_\alpha\) and smaller MAE).
Deep learning models that better models are shown in the upper left corner (i.e., with a larger $S_\alpha$ and smaller MAE). From Fig. 5, we have following observations:

**Traditional vs. Deep Models.** Compared with traditional RGB-D based SOD models, deep learning models obtain significantly better performance. This confirms the powerful feature...
Fig. 7 F-measures under different thresholds for 24 RGB-D based models on STERE [96], NLPR [99], LFSD [75], DES [14], SIP [34], GIT [16], SSD [189], and NJUD [63] datasets.

Tab. 9 Attribute-based study w.r.t. background clutter. Comparison results for 24 representative RGB-D based SOD models (9 traditional models and 15 deep learning models) are provided in terms of MAE and $S_a$. The three best results are shown in red, blue and green fonts.

| Traditional models | Deep learning models |
|--------------------|----------------------|
| **MAE**            | **S**                |
| Simple             | 1.23 0.86 0.48      | 1.23 0.86 0.48 |
| Uncertain          | 1.52 1.07 0.69      | 1.52 1.07 0.69 |
| Complex            | 1.86 1.39 0.98      | 1.86 1.39 0.98 |

- **Comparison of Deep Models.** Among the deep learning-based models, D3Net [34], JL-DCF [12], UC-Net [160], SSF [167], ICNet [71], and S2MA [82] obtain better performance than other deep models.

Moreover, Fig. 6 and Fig. 7 show the PR and F-measure curves of the 24 representative RGB-D based...
Deep learning models of images have small, medium, and big salient objects collected from STERE [96], NLPR [99], LFSD [75], evaluation, we build a hybrid dataset with 2464 images the range of \([0, 0.4]\), it is denoted as “big”; and 3) when the ratio is in it is denoted as “small”; 2) when the ratio is larger than this ratio, we compute the ratio between the size of object area, we compute the ratio between the size of RGB-D based SOD models. based evaluations on the performance of representative and lighting conditions, we carry out diverse attributebased study and five other methods, namely UC-Net [160], CPFP [173], S2MA [82], ICNet [71], JL-DCF [42], and NJUD [63]. Note that, there are 1000, 300, 100, 500, 200, and 80 samples as testing for the NLPR, LFSD, DES, SIP, GIT, and SSD, respectively. For the NJUD [63] dataset, there are 485 images as testing for CPFP [173], S2MA [82], ICNet [71], JL-DCF [42], and UC-Net [160], while 498 testing images for all other models.

### 5.2.2 Attribute-based Evaluation

To investigate the influence of different factors, such as object scale, background clutter, numbers of salient objects, indoor or outdoor scene, background objects, and lighting conditions, we carry out diverse attribute-based evaluations on the performance of representative RGB-D based SOD models.

- **Object Scale.** To characterize the scale of a salient object area, we calculate the ratio between the size of the salient area and the whole image. We define three types of object scales: 1) when the ratio is less than 0.1, it is denoted as “small”; 2) when the ratio is larger than 0.4, it is denoted as “big”; and 3) when the ratio is in the range of \([0.1, 0.4]\), it is denoted as “medium”. In this evaluation, we build a hybrid dataset with 2464 images collected from STERE [96], NLPR [99], LFSD [75], DES [14], and SIP [34], where 24%, 69.2% and 6.8% of images have small, medium, and big salient object areas, respectively. The constructed hybrid dataset can be found at https://github.com/taozh2017/RGBD-SOdsurvey. The comparison results of the attribute-based study \(w.r.t.\) object scale are shown in Tab. 8.

![Fig. 8 Sampling images from three types of background cluster.](https://github.com/taozh2017/RGBD-SOdsurvey)
Fig. 9  Attribute-based study w.r.t. number of salient object(s) (i.e., single vs. multiple (multi)). The comparison results on 24 representative RGB-D based SOD models (i.e., LHM [99], ACSD [63], DESM [14], GP [111], LBE [41], DCMC [22], SE [44], CDCP [19], CDB [79], DF [109], PCF [4], CTMF [46], CPFP [173], TANet [6], AFNet [132], MMCI [9], DMRA [101], D3Net [34], SSF [167], A2dele [106], S2MA [82], ICNet [71], JL-DCF [42], and UC-Net [160]) in terms of MAE (top) and $S_\alpha$ (bottom) metrics.

JL-DCF [42], UC-Net [160], and S2MA [82], obtain the best performance. D3Net [34], SSF [167], A2dele [106], and ICNet [71] also obtain promising performance.

- **Background Clutter.** It is difficult to directly characterize background clutter. Since classic SOD methods tend to use prior information or color contrast to locate salient objects, they often fail under complex backgrounds. Thus, in this evaluation, we utilize five traditional SOD methods, i.e., BSCA [108], CLC [179], MDC [56], MIL [51], and WFD [55], to first detect salient objects in various images and then group these images into different categories (e.g., simple or complex background) according to the results. Specifically, we first construct a hybrid dataset with 1400 images collected from three datasets (STERE [96], NLPR [99], and LFSD [75]). Then, we conduct the five models on this dataset and obtain the $S_\alpha$ values for each, which we use to characterize images as follows: 1) If all $S_\alpha$ values are more than 0.9, the image is denoted as having a “simple” background; 2) If all $S_\alpha$ values are less than 0.6, the image is said to have a “complex” background; 3) The remaining images are denoted as “uncertain”. Some example images with the three types of background clutter are shown in Fig. 8. The constructed hybrid dataset can be found at https://github.com/taozh2017/RGBD-SODsurvey. The comparison results of attribute-based study w.r.t. background clutter are shown in Tab. 9. As can be seen from the results, all models obtain worse SOD performance on images containing complex backgrounds than simple ones. Among the representative models, JL-DCF [42], UC-Net [160] and SSF [167] achieve top-three best results. Besides, the four most recent models, i.e., D3Net [34], S2MA [82], A2dele [106], and ICNet [71] also obtain relatively better performance than the other models.

- **Single vs. Multiple Objects.** In this evaluation, we construct a hybrid dataset with 1229 images collected from the NLPR [99] and SIP [34] datasets. The comparison results are shown in Fig. 9. From
the results, we can see that it is easier to detect single salient object than multiple ones.

- **Indoor vs. Outdoor.** We evaluate the performance of different RGB-D based SOD models on indoor and outdoor scenes. In this evaluation, we construct a hybrid dataset collected from the DES [14], NLPR [99], and LFSD [75] datasets. The comparison results of attribute-based study w.r.t. indoor vs. outdoor are shown in Fig. 10. From the results, it can be seen that most models difficulty detect salient objects on indoor scene than outdoor ones. This is possibly because indoor environments often suffer from uncertain light conditions.

- **Background Objects.** We evaluate the...
Performance of the RGB-D based SOD models when different background objects are present. We use SIP dataset [34], and split it into nine categories, i.e., car, barrier, flower, grass, road, sign, tree, and other. The comparison results are shown in Tab. 10. As can be seen, all methods obtain diverse performances under different background objects. Among the 24 representative RGB-D based models, JL-DCF [42], UC-Net [160] and SSF [167] achieve the top-three best results. In addition, the four most recent models, i.e., D3Net [34], S2MA [82], A2dele [106], and ICNet [71] obtain relatively better performance than the others.

- Lighting Conditions. The performance of SOD can be affected by different lighting conditions. To determine the performance of different RGB-D based SOD models under different lighting conditions, we conduct an evaluation on the SIP dataset [34], which we split it into two categories, i.e., sunny and low-light. The comparison results are shown in Tab. 11. As can be seen, low-light negatively impacts SOD performance. Specifically, UC-Net [160] obtains the best performance under sunny condition while JL-DCF [42] achieves the best result under low-light condition.

In addition, we report the saliency maps generated for various challenging scenes to visualize the performance of different RGB-D based SOD models. Fig. 11 and Fig. 12 show some representative examples using two classic non-deep methods (DCMC [22] and

---

**Fig. 11** Visual comparisons for two classical non-deep methods (DCMC [22] and SE [44]) and three state-of-the-art CNN-based models (DMRA [101], D3Net [34], SSF [167]).
Fig. 12 Visual comparisons for five state-of-the-art CNN-based models (A2dele [106], S2MA [82], ICNet [71], JL-DCF [42], and UC-Net [160]).

SE [44]) and eight state-of-the-art CNN-based models (DMRA [101], D3Net [34], SSF [167], A2dele [106], S2MA [82], ICNet [71], JL-DCF [42], and UC-Net [160]). The 1st row shows a small object, while 2nd row is an example of a big one. The 3rd row and 4th rows contain complex background and boundaries, respectively. The 5th and 6th rows contain multiple salient objects. In the 7th row, there is low-light condition. In the 8th row, the depth map is coarse with very inaccurate object boundaries, which could inhibit the SOD performance. From the results in Fig. 11 and Fig. 12, it can be observed that deep models perform better than non-deep models on these challenging scenes, confirming the powerful expression ability of deep features over handcrafted ones. In addition, D3Net [34], S2MA [82], JL-DCF [42], and UC-Net [160] perform better than other deep models.

6 Challenges and Open Directions

6.1 Effects of Imperfect Depth

Effects of Low-quality Depth Maps. Depth maps with affluent spatial information have been proven beneficial in detecting salient objects from cluttered backgrounds, while the depth quality directly affects the subsequent SOD performance. The quality of depth maps varies tremendously across different scenarios due to limitations of depth Sensors, posing
a challenge when trying to reduce the effects of low-quality depth maps. However, most existing methods directly fuse RGB images and original raw data from depth maps, without considering the effects of low-quality depth maps. There are a few notable exceptions. For example, in [173], a contrast enhanced network was proposed to learn enhanced depth maps, which have much higher contrasts compared with the original depths. In [167], a compensation-aware loss was designed to pay more attention to some hard samples containing unreliable depth information. Moreover, $D^3$Net [34] uses a depth deuptator unit (DDU) to classify depth maps into two classes (i.e., reasonable and low-quality). The DDU also acts as a gate that can filter out the low-quality depth maps. However, the above methods often employ a two-step strategy to achieve depth enhancement and multi-modal fusion [167, 173] or an independent gate operation for filtering out poor depths, which could bring a suboptimal problem. There is thus a need to develop an end-to-end framework that can achieve depth enhancement or adaptively weight the depth maps (e.g., assigns low weights for poor depth maps) during multi-modal fusion, which would be more helpful for reducing the effects of low-quality depth maps and boosting SOD performance.

**Incomplete Depth Maps.** In RGB-D datasets, it is inevitable for there to be some low-quality depth maps due to the limitations of the acquisition devices. As previously discussed, several depth enhancement algorithms have been used to improve the quality of depth maps. However, when some depth maps suffer from severe noise or blurred edges, these depth maps could be discarded. In this case, we have complete RGB images but some samples without having depth maps, which is similar to the incomplete multi-view/modal learning problem [147, 180–183]. Thus, we call it “incomplete RGB-D based SOD”. As current models only focus on the SOD task using complete RGB images and depth maps, we believe this could be a new direction for RGB-D SOD.

**Depth Estimation.** Furthermore, depth estimation provides an effective solution to recovery high-quality depths and overcome the effects of low-quality depth maps. Various depth estimation approaches [43, 61, 81, 131] have been developed, which could be introduced into the RGB-D based SOD task to improve performance.

### 6.2 Effective Fusion Strategies

**Adversarial Learning-based Fusion.** It is important to effectively fuse RGB images and depth maps for RGB-D based SOD. Existing models often employ different fusion strategies (e.g., early fusion, middle fusion, or late fusion) to exploit the correlations between RGB images and depth maps. Recently, generative adversarial networks (GANs) [93] have gained widespread attention for the salient detection task [97, 191]. In common GAN-based SOD models, a generator takes RGB images as inputs and generates the corresponding saliency maps, while a discriminator is adopted to distinguish whether a given image is synthetic or ground-truth. GAN-based model could easily be extended to RGB-D SOD, which could be helpful for boosting performance due to their superior feature learning ability. Moreover, GANs could also be used to learn the common feature representations for RGB images and depth maps [59], which could help with feature or saliency map fusion and further boost the SOD performance.

**Attention-induced Fusion.** Attention mechanisms have been widely applied to various deep learning-based tasks [39, 126, 129, 139], allowing networks to selectively pay attention to a subset of regions for extracting discriminative and powerful features. Besides, co-attention mechanisms have been developed to explore the underlying correlations across multiple modalities, and are widely studied in visual question answering [86, 152] and video object segmentation [87]. Thus, for RGB-D based SOD task, we could also develop attention-based fusion algorithms to exploit correlations between RGB images and depth cues to improve the performance.

### 6.3 Different Supervision Strategies

Existing RGB-D models often used a fully supervised strategy to learn saliency prediction models. However, annotating pixel-level saliency maps is a tedious and time-consuming procedure. To alleviate this issue, there has been increased interest in weakly and semi-supervised learning, which have been applied to salient object detection [107, 149, 154, 158, 187]. Semi-/weak supervision could also be introduced into RGB-D SOD, by leveraging image-level tags [154] and pseudo pixel-wise annotations [149, 156], for improving the detection performance. Besides, several studies [12, 23] have suggested that models pretrained using self-supervision can effectively be used for achieving better performance. Therefore, we could train saliency prediction models on large amounts of annotated RGB
images in a self-supervised manner and then transfer the pretrained models to the RGB-D SOD task.

6.4 Dataset Collection

Large-scale. Although there are nine public RGB-D datasets for SOD, their size is quite limited, e.g., the maximum size is about 2000 samples for NJUD [63]. When compared with other RGB-D datasets for generic object detection or action recognition [64, 161], the size of RGB-D datasets for SOD is also very small. Thus, it is essential to develop new large-scale RGB-D datasets that can serve as baselines for future research.

Complex Background & Task-driven. Most existing RGB-D datasets collect images that contain one salient object or multiple objects but with a relatively clean background. However, real-world applications often suffer from much more complicated situations (e.g., occlusion, appearance change, low illumination, etc), which could decrease the SOD performance. Thus, collecting images with complex background is critical to improve the generalization ability of RGB-D SOD models. Moreover, for some tasks, images with specific salient object(s) must be collected. For example, one important technology is road sign recognition in driver assistance systems, which requires images with road signs to be collected. Thus, it is essential to construct task-driven RGB-D datasets like SIP [34].

6.5 Model Design for Real-world Scenarios

Some smartphones can capture depth maps (e.g., images in the SIP dataset were captured using Huawei Mate 10). Thus it would be feasible to conduct the SOD task in real-world applications, e.g., on smart devices. However, most existing methods include complicated and deep DNNs to increase the model capacity and achieve better performance, preventing them from being directly applied to real-work platforms. To overcome this, model compression [15, 48] techniques could be used to learn compact RGB-D based SOD models with promising detection accuracy. Moreover, JL-DFC utilizes a shared network to locate salient objects using RGB and depth views, which largely reduces the model parameters and makes real-world applications feasible.

6.6 Extension to RGB-T SOD

In addition to RGB-D SOD, there are several other methods fusing different modalities for better detection, such as RGB-T SOD, which integrates RGB and thermal infrared data. Thermal infrared cameras can capture the radiation emitted from any object with a temperature above absolute zeros, making thermal infrared images insensitive to illumination conditions [90]. Therefore, thermal images can provide supplementary information to improve SOD performance when salient objects suffer from varying light, reflective light, or shadows. Some RGB-T models [69, 90, 120, 122–125, 130, 170] and datasets (VT821 [130], VT1000 [125] and VT5000 [123]) have already been proposed over the past few years. Similar to RGB-D SOD, the key aim of RGB-T SOD is to fuse RGB and thermal infrared images and exploit the correlations between the two modalities. Thus, several advanced multi-modal fusion technologies in RGB-D SOD could be extended to the RGB-T SOD task.

7 Conclusion

In this paper, to the best of our knowledge, we present the first comprehensive review of RGB-D based SOD models. We first review the models from different perspectives, and then summarize popular RGB-D SOD datasets as well as provide details for each. Considering that the light field also provides depth information, we also review popular light field SOD models and the related benchmark datasets. Next, we provide a comprehensive evaluation of 24 representative RGB-D based SOD models as well as an attribute-based evaluation. Specifically, we perform attribute-based performance analysis by constructing new datasets for the 24 representative RGB-D based SOD models. Moreover, we discuss several challenges and highlight open directions for future research. In addition, we briefly discuss the extension work to RGB-T SOD to improve performance when salient objects suffer from varying light, reflective light, or shadows. Although RGB-D based SOD has made notable progress in the past several decades, there is still significant room for improvement. We hope this survey will generate more interest works in RGB-D based SOD.

Acknowledgements

This research was supported by Major Project for New Generation of AI under Grant No. 2018AAA0100400, NSFC (61922046), and Tianjin Natural Science Foundation (17JCJQJC43700).

Open Access

This article is distributed under the terms of the Creative Commons Attribution License which permits any use, distribution, and reproduction in any medium, provided the original author(s) and the source are credited.
References

[1] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, and J. Li. Salient object detection: A survey. Computational Visual Media, pages 1–34, 2014.

[2] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. Salient object detection: A benchmark. IEEE Transactions on Image Processing, 24(12):5706–5722, 2015.

[3] C. Chen, J. Wei, C. Peng, W. Zhang, and H. Qin. Improved saliency detection in RGB-D images using two-phase depth estimation and selective deep fusion. IEEE Transactions on Image Processing, 29:4296–4307, 2020.

[4] H. Chen and Y. Li. Progressively complementarity-aware fusion network for RGB-D salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3051–3060, 2018.

[5] H. Chen and Y. Li. CNN-based rgb-d salient object detection: Learn, select and fuse. arXiv preprint arXiv:1909.09309, 2019.

[6] H. Chen and Y. Li. Three-stream attention-aware network for RGB-D salient object detection. IEEE Transactions on Image Processing, 28(6):2825–2835, 2019.

[7] H. Chen, Y. Li, and D. Su. RGB-D saliency detection by multi-stream late fusion network. In Proceedings of the International Conference on Computer Vision Systems, pages 459–468. Springer, 2017.

[8] H. Chen, Y. Li, and D. Su. Discriminative cross-modal transfer learning and densely cross-level feedback fusion for RGB-D salient object detection. IEEE Transactions on Cybernetics, 2019.

[9] H. Chen, Y. Li, and D. Su. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. Pattern Recognition, 86:376–385, 2019.

[10] H. Chen, Y.-F. Li, and D. Su. M3net: Multi-scale multi-path multi-modal fusion network and example application to rgb-d salient object detection. In Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 4911–4916. IEEE, 2017.

[11] H. Chen, Y.-F. Li, and D. Su. Attention-aware cross-modal cross-level fusion network for RGB-D salient object detection. In Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems, pages 6821–6826. IEEE, 2018.

[12] T. Chen, S. Liu, S. Chang, Y. Cheng, L. Amini, and Z. Wang. Adversarial robustness: From self-supervised pre-training to fine-tuning. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 699–708, 2020.

[13] Z. Chen and Q. Huang. Depth potentiality-aware gated attention network for RGB-D salient object detection. arXiv preprint arXiv:2003.08608, 2020.

[14] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao. Depth enhanced saliency detection method. In Proceedings of the International Conference on Internet Multimedia Computing and Service, pages 23–27, 2014.

[15] Y. Cheng, D. Wang, P. Zhou, and T. Zhang. A survey of model compression and acceleration for deep neural networks. arXiv preprint arXiv:1710.09282, 2017.

[16] A. Ciptadi, T. Herrmans, and J. M. Rehg. An in depth view of saliency. Georgia Institute of Technology, 2013.

[17] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang. Review of visual saliency detection with comprehensive information. IEEE Transactions on Circuits and Systems for Video Technology, 29(10):2941–2959, 2018.

[18] R. Cong, J. Lei, H. Fu, J. Hou, Q. Huang, and S. Kwong. Going from RGB to RGBD saliency: A depth-guided transformation model. IEEE Transactions on Cybernetics, 2019.

[19] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and C. Hou. Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation. IEEE Transactions on Image Processing, 27(2):568–579, 2017.

[20] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and N. Ling. HSCS: Hierarchical sparsity based co-saliency detection for RGBD images. IEEE Transactions on Multimedia, 21(7):1660–1671, 2018.

[21] R. Cong, J. Lei, H. Fu, W. Lin, Q. Huang, X. Cao, and C. Hou. An iterative co-saliency framework for RGBD images. IEEE Transactions on Cybernetics, 49(1):233–246, 2017.

[22] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, and C. Hou. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. IEEE Signal Processing Letters, 23(6):819–823, 2016.

[23] A. Dai, C. Diller, and M. Nießner. Sg-nn: Sparse generative neural networks for self-supervised scene completion of rgb-d scans. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 849–858, 2020.

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

[25] Y. Ding, Z. Liu, M. Huang, R. Shi, and X. Wang. Depth-aware saliency detection using convolutional neural networks. Journal of Visual Communication and Image Representation, 61:1–9, 2019.

[26] H. Du, Z. Liu, and R. Shi. Saliency object segmentation based on depth-aware image layering. Multimedia Tools and Applications, 78(9):12125–12138, 2019.

[27] H. Du, Z. Liu, H. Song, L. Mei, and Z. Xu. Improving RGBD saliency detection using progressive region classification and saliency fusion. IEEE Access, 4:8987–8994, 2016.

[28] D.-P. Fan, M.-M. Cheng, J.-J. Liu, S.-H. Gao, Q. Hou, and A. Borji. Saliency objects in clutter:
Bringing salient object detection to the foreground. In *Proceedings of the European Conference on Computer Vision*, pages 186–202. Springer, 2018.

[29] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji. Structure-measure: A new way to evaluate foreground maps. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4548–4557, 2017.

[30] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji. Enhanced-alignment measure for binary foreground map evaluation. In *Proceedings of the International Joint Conferences on Artificial Intelligence*, pages 698–704, 2018.

[31] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao. Camouflaged object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2777–2787, 2020.

[32] D.-P. Fan, G.-P. Ji, T. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao. Pranet: Parallel reverse attention network for polyp segmentation. In *Medical Image Computing and Computer-Assisted Intervention*, 2020.

[33] D.-P. Fan, T. Li, Z. Lin, G.-P. Ji, D. Zhang, M.-M. Cheng, H. Fu, and J. Shen. Re-thinking co-salient object detection. *arXiv preprint arXiv:2007.03380*, 2020.

[34] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[35] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen. Shifting more attention to video salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8554–8564, 2019.

[36] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, and L. Shao. Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network. In *Proceedings of the European Conference on Computer Vision*. Springer, 2020.

[37] D.-P. Fan, T. Zhou, G.-P. Ji, Y. Zhou, G. Chen, H. Fu, J. Shen, and L. Shao. Inf-net: Automatic covid-19 lung infection segmentation from ct images. *IEEE Transactions on Medical Imaging*, 2020.

[38] X. Fan, Z. Liu, and G. Sun. Salient region detection for stereoscopic images. In *Proceedings of the International Conference on Digital Signal Processing*, pages 454–458. IEEE, 2014.

[39] H.-S. Fang, J. Cao, Y.-W. Tai, and C. Lu. Pairwise body-part attention for recognizing human-object interactions. In *Proceedings of the European Conference on Computer Vision*, pages 51–67. Springer, 2018.

[40] D. Feng, N. Barnes, and S. You. Hosu: Histogram of surface orientation for rgb-d salient object detection. In *Proceedings of the International Conference on Digital Image Computing: Techniques and Applications*, pages 1–8. IEEE, 2017.

[41] D. Feng, N. Barnes, S. You, and C. McCarthy. Local background enclosure for RGB-D salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2343–2350, 2016.

[42] K. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao. Jl-dcf: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2020.

[43] C. Godard, O. Mac Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 270–279, 2017.

[44] J. Guo, T. Ren, and J. Bei. Salient object detection for RGB-D image via saliency evolution. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 1–6. IEEE, 2016.

[45] J. Guo, T. Ren, J. Bei, and Y. Zhu. Salient object detection in rgb-d image based on saliency fusion and propagation. In *Proceedings of the International Conference on Internet Multimedia Computing and Service*, pages 1–5, 2015.

[46] J. Han, H. Chen, N. Liu, C. Yan, and X. Li. Cnns-based RGB-D saliency detection via cross-view transfer and multiview fusion. *IEEE Transactions on Cybernetics*, 48(11):3171–3183, 2017.

[47] J. Han, D. Zhang, G. Cheng, N. Liu, and D. Xu. Advanced deep-learning techniques for salient and category-specific object detection: a survey. *IEEE Signal Processing Magazine*, 35(1):84–100, 2018.

[48] Y. He, J. Lin, Z. Liu, H. Wang, L.-J. Li, and S. Han. Amc: Automl for model compression and acceleration on mobile devices. In *Proceedings of the European Conference on Computer Vision*, pages 784–800. Springer, 2018.

[49] S. Hong, T. You, S. Kwak, and B. Han. Online tracking by learning discriminative saliency map with convolutional neural network. In *Proceedings of the International Conference on Machine Learning*, pages 597–606, 2015.

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

[51] F. Huang, J. Qi, H. Lu, L. Zhang, and X. Ruan. Salient object detection via multiple instance learning. *IEEE Transactions on Image Processing*, 26(4):1911–1922, 2017.

[52] P. Huang, C.-H. Shen, and H.-F. Hsiao. Rgbd salient object detection using spatially coherent deep learning framework. In *Proceedings of the IEEE International Conference on Digital Signal Processing*, pages 1–5.
IEEE, 2018.

[53] R. Huang, Y. Xing, and Z. Wang. RGB-D salient object detection by a CNN with multiple layers fusion. *IEEE Signal Processing Letters*, 26(4):552–556, 2019.

[54] R. Huang, Y. Xing, and Y. Zou. Triple-complementary network for RGB-D salient object detection. *IEEE Signal Processing Letters*, 2020.

[55] X. Huang and Y. Zhang. Water flow driven salient object detection at 180 fps. *Pattern Recognition*, 76:95–107, 2018.

[56] X. Huang and Y.-J. Zhang. 300-fps salient object detection by a CNN with multiple layers fusion. *IEEE Signal Processing Letters*, 26(9):4243–4254, 2017.

[57] N. Imamoglu, W. Shimoda, C. Zhang, Y. Fang, A. Kanazaki, K. Yanai, and Y. Nishida. An integration of bottom-up and top-down salient cues on rgb-d data: saliency from objectness versus non-objectness. *Signal, Image and Video Processing*, 12(2):307–314, 2018.

[58] W. Ji, J. Li, M. Zhang, Y. Piao, and H. Lu. Accurate rgb-d salient object detection via collaborative learning. In *ECCV*, 2020.

[59] B. Jiang, Z. Zhou, X. Wang, J. Tang, and B. Luo. cmsalgan: RGB-D salient object detection with cross-view generative adversarial networks. *IEEE Transactions on Multimedia*, 2020.

[60] L. Jiang, A. Koch, and A. Zell. Salient regions detection for indoor robots using rgb-d data. In *IEEE International Conference on Robotics and Automation*, pages 1323–1328. IEEE, 2015.

[61] L. Jin, Y. Xu, J. Zheng, J. Zhang, R. Tang, S. Xu, J. Yu, and S. Gao. Geometric structure based and regularized depth estimation from 360 indoor imagery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 889–898, 2020.

[62] Z. Jin, J. Li, and D. Li. Co-saliency detection for rgbd images based on effective propagation mechanism. *IEEE Access*, 7:141311–141318, 2019.

[63] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu. Depth saliency based on anisotropic center-surround difference. In *Proceedings of the IEEE International Conference on Image Processing*, pages 1115–1119. IEEE, 2014.

[64] K. Lai, L. Bo, X. Ren, and D. Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *Proceedings of the IEEE International Conference on Robotics and Automation*, pages 1817–1824. IEEE, 2011.

[65] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanahalli, and S. Yan. Depth matters: Influence of depth cues on visual saliency. In *Proceedings of the European Conference on Computer Vision*, pages 101–115. Springer, 2012.

[66] J. Lei, H. Zhang, L. You, C. Hou, and L. Wang. Evaluation and modeling of depth feature incorporated visual attention for salient object segmentation. *Neurocomputing*, 120:24–33, 2013.

[67] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, and Q. Huang. ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection. *IEEE Transactions on Cybernetics*, 2020.

[68] C. Li, R. Cong, Y. Piao, Q. Xu, and C. C. Loy. Rgb-d salient object detection with cross-modality modulation and selection. In *Proceedings of the European Conference on Computer Vision*. Springer, 2020.

[69] C. Li, G. Wang, Y. Ma, A. Zheng, B. Luo, and J. Tang. A unified rgb-t saliency detection benchmark: dataset, baselines, analysis and a novel approach. *arXiv preprint arXiv:1701.02829*, 2017.

[70] C. Li, B. Zhan, S. Zhang, and H. Sheng. Saliency detection with relative location measure in light field image. In *Proceedings of the International Conference on Image, Vision and Computing*, pages 8–12. IEEE, 2017.

[71] G. Li, Z. Liu, and H. Ling. Icnet: Information conversion network for RGB-D based salient object detection. *IEEE Transactions on Image Processing*, 29:4873–4884, 2020.

[72] G. Li, Z. Liu, L. Ye, Y. Wang, and H. Ling. Cross-modal weighting network for rgb-d salient object detection. In *Proceedings of the European Conference on Computer Vision*. Springer, 2020.

[73] G. Li and Y. Yu. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 478–487, 2016.

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

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

[76] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu. Saliency detection on light field. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(8):1605–1616, 2017.

[77] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen. Contour knowledge transfer for salient object detection. In *Proceedings of the Proceedings of the European Conference on Computer Vision*. Springer, September 2018.

[78] F. Liang, L. Duan, W. Ma, Y. Qiao, Z. Cai, J. Miao, and Q. Ye. Cognn: RGB-D deep fusion for stereoscopic salient object detection. *Pattern Recognition*, page 107329, 2020.

[79] F. Liang, L. Duan, W. Ma, Y. Qiao, Z. Cai, and L. Qing. Stereoscopic saliency model using contrast and depth-guided-background prior. *Neurocomputing*, 275:2227–2238, 2018.

[80] D. Liu, Y. Hu, K. Zhang, and Z. Chen. Two-stream
refinement network for rgb-d saliency detection. In Proceedings of IEEE International Conference on Image Processing, pages 3925–3929. IEEE, 2019.

[81] F. Liu, C. Shen, and G. Lin. Deep convolutional neural fields for depth estimation from a single image. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 5162–5170, 2015.

[82] N. Liu, N. Zhang, and J. Han. Learning selective self-mutual attention for RGB-D saliency detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[83] Z. Liu, S. Shi, Q. Duan, W. Zhang, and P. Zhao. Saliency object detection for RGB-D image by single stream recurrent convolution neural network. Neurocomputing, 363:46–57, 2019.

[84] Z. Liu, J. Tang, Q. Xiang, and P. Zhao. Saliency object detection for rgb-d images by generative adversarial network. Multimedia Tools and Applications, pages 1–23, 2020.

[85] Z. Liu, W. Zhang, and P. Zhao. A cross-modal adaptive gated fusion generative adversarial network for RGB-D salient object detection. Neurocomputing, 2020.

[86] J. Lu, J. Yang, D. Batra, and D. Parikh. Hierarchical question-image co-attention for visual question answering. In Advances in neural information processing systems, pages 289–297, 2016.

[87] X. Lu, W. Wang, C. Ma, J. Shen, L. Shao, and F. Porikli. See more, know more: Unsupervised video object segmentation with co-attention siamese networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3623–3632, 2019.

[88] A. Luo, X. Li, F. Yang, Z. Jiao, H. Cheng, and S. Lyu. Cascade graph neural networks for rgb-d salient object detection. In Proceedings of the Proceedings of the European Conference on Computer Vision. Springer, 2020.

[89] C.-Y. Ma and H.-M. Hang. Learning-based saliency model with depth information. Journal of vision, 15(6):19–19, 2015.

[90] Y. Ma, D. Sun, Q. Meng, Z. Ding, and C. Li. Learning multiscale deep features and svm regressors for adaptive rgb-t saliency detection. In Proceedings of the International Symposium on Computational Intelligence and Design, volume 1, pages 389–392. IEEE, 2017.

[91] V. Mahadevan and N. Vasconcelos. Saliency-based discriminant tracking. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1007–1013. IEEE, 2009.

[92] N. Martine, C. Micheloni, and G. L. Foresti. Kernelized saliency-based person re-identification through multiple metric learning. IEEE Transactions on Image Processing, 24(12):5645–5658, 2015.

[93] M. Mirza and S. Osindero. Conditional generative adversarial nets. arXiv preprint arXiv:1411.1784, 2014.

[94] T. V. Nguyen, Q. Zhao, and S. Yan. Attentive systems: A survey. International Journal of Computer Vision, 126(1):86–110, 2018.

[95] G.-Y. Nie, M.-M. Cheng, Y. Liu, Z. Liang, D.-P. Fan, Y. Liu, and Y. Wang. Multi-level context ultrasound-aggregation for stereo matching. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3283–3291, 2019.

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

[97] J. Pan, C. C. Ferrer, K. McGuinness, et al. Salgan: Visual saliency prediction with generative adversarial networks. arXiv preprint arXiv:1701.01081, 2017.

[98] Y. Pang, L. Zhang, X. Zhao, and H. Lu. Hierarchical dynamic filtering network for RGB-D salient object detection. In Proceedings of the European Conference on Computer Vision. Springer, 2020.

[99] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji. Rgbd salient object detection: a benchmark and algorithms. In Proceedings of the European Conference on Computer Vision, pages 92–109. Springer, 2014.

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

[101] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 7254–7263, 2019.

[102] Y. Piao, X. Li, and M. Zhang. Depth-induced cellular automata for light field saliency. In Frontiers in Optics, pages FTh3E–3. Optical Society of America, 2018.

[103] Y. Piao, X. Li, M. Zhang, J. Yu, and H. Lu. Saliency detection via depth-induced cellular automata on light field. IEEE Transactions on Image Processing, 29:1879–1889, 2019.

[104] Y. Piao, Z. Rong, M. Zhang, X. Li, and H. Lu. Deep light-field-driven saliency detection from a single view. In Proceedings of the International Joint Conference on Artificial Intelligence, 2019.

[105] Y. Piao, Z. Rong, M. Zhang, and H. Lu. Exploit and replace: An asymmetrical two-stream architecture for versatile light field saliency detection. In Proceedings of the Association for the Advancement of Artificial Intelligence, 2020.

[106] Y. Piao, Z. Rong, M. Zhang, W. Ren, and H. Lu. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[107] M. Qian, J. Qi, L. Zhang, M. Feng, and H. Lu.
Language-aware weak supervision for salient object detection. Pattern Recognition, 96:106955, 2019.

[108] Y. Qin, H. Lu, Y. Xu, and H. Wang. Salient detection via cellular automata. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 110–119, 2015.

[109] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang. RGBD salient object detection via deep fusion. IEEE Transactions on Image Processing, 26(5):2274–2285, 2017.

[110] K. Rapantzikos, Y. Avrithis, and S. Kollias. Dense saliency-based spatiotemporal feature points for action recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 1454–1461. IEEE, 2009.

[111] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Ying Yang. Exploiting global priors for RGB-D saliency detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pages 25–32, 2015.

[112] H. Sheng, X. Liu, and S. Zhang. Saliency analysis based on depth contrast increased. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, pages 1347–1351. IEEE, 2016.

[113] H. Sheng, S. Zhang, X. Liu, and Z. Xiong. Relative location for light field saliency detection. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1631–1635. IEEE, 2016.

[114] R. Shigematsu, D. Feng, S. You, and N. Barnes. Learning RGB-D salient object detection using background enclosure, depth contrast, and top-down features. In Proceedings of the IEEE International Conference on Computer Vision Workshops, pages 2749–2757, 2017.

[115] W. Shimoda and K. Yanai. Distinct class-specific saliency maps for weakly supervised semantic segmentation. In Proceedings of the European Conference on Computer Vision, pages 218–234. Springer, 2016.

[116] H. Song, Z. Liu, H. Du, and G. Sun. Depth-aware saliency detection using discriminative saliency fusion. In Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pages 1626–1630. IEEE, 2016.

[117] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, and T. Ren. Depth-aware salient object detection and segmentation via multiscalar discriminative saliency fusion and bootstrap learning. IEEE Transactions on Image Processing, 26(9):4204–4216, 2017.

[118] H. Song, W. Wang, S. Zhao, J. Shen, and K.-M. Lam. Pyramid dilated deeper convlstm for video salient object detection. In Proceedings of the European Conference on Computer Vision, pages 715–731. Springer, 2018.

[119] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian. Selectivity or invariance: Boundary-aware salient object detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 3799–3808, 2019.

[120] D. Sun, S. Li, Z. Ding, and B. Luo. Rgb-t saliency detection via robust graph learning and collaborative manifold ranking. In Proceedings of the International Conference on Bio-Inspired Computing: Theories and Applications, pages 670–684. Springer, 2019.

[121] Y. Tang, R. Tong, M. Tang, and Y. Zhang. Depth incorporating with color improves salient object detection. The Visual Computer, 32(1):111–121, 2016.

[122] Z. Tu, Z. Li, C. Li, Y. Lang, and J. Tang. Multi-interactive encoder-decoder network for rgbt salient object detection. arXiv preprint arXiv:2005.02315, 2020.

[123] Z. Tu, Y. Ma, Z. Li, C. Li, J. Xu, and Y. Liu. Rgbt salient object detection: A large-scale dataset and benchmark. arXiv preprint arXiv:2007.03262, 2020.

[124] Z. Tu, T. Xia, C. Li, Y. Lu, and J. Tang. M3s-nir: Multi-modal multi-scale noise-insensitive ranking for rgb-t saliency detection. In Proceedings of the IEEE Conference on Multimedia Information Processing and Retrieval, pages 141–146. IEEE, 2019.

[125] Z. Tu, T. Xia, C. Li, X. Wang, Y. Ma, and J. Tang. Rgb-t image saliency detection via collaborative graph learning. IEEE Transactions on Multimedia, 22(1):160–173, 2019.

[126] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In Proceedings of the conference on Neural Information Processing Systems, pages 5998–6008, 2017.

[127] A. Wang and M. Wang. RGB-D salient object detection via minimum barrier distance transform and saliency fusion. IEEE Signal Processing Letters, 24(5):663–667, 2017.

[128] A. Wang, M. Wang, X. Li, Z. Mi, and H. Zhou. A two-stage bayesian integration framework for salient object detection on light field. Neural Processing Letters, 46(3):1083–1094, 2017.

[129] F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, and X. Tang. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3156–3164. Springer, 2017.

[130] G. Wang, C. Li, Y. Ma, A. Zheng, J. Tang, and B. Luo. Rgb-t saliency detection benchmark: Dataset, baselines, analysis and a novel approach. In Proceedings of the Chinese Conference on Image and Graphics Technologies, pages 359–369. Springer, 2018.

[131] L. Wang, J. Zhang, O. Wang, Z. Lin, and H. Lu. Sde-depth: Semantic divide-and-conquer network for monocular depth estimation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 541–550, 2020.

[132] N. Wang and X. Gong. Adaptive fusion for RGB-D
salient object detection. *IEEE Access*, 7:55277–55284, 2019.

133. S. Wang, W. Liao, P. Surman, Z. Tu, Y. Zheng, and J. Yuan. Salience guided depth calibration for perceptually optimized compressive light field 3d display. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2031–2040, 2018.

134. S.-T. Wang, Z. Zhou, H.-B. Qu, and B. Li. Rgb-d saliency detection under bayesian framework. In *Proceedings of International Conference on Pattern Recognition*, pages 1881–1886. IEEE, 2016.

135. S.-T. Wang, Z. Zhou, H.-B. Qu, and B. Li. Visual saliency detection for RGB-D images with generative model. In *Proceedings of the Asian Conference on Computer Vision*, pages 20–35. Springer, 2016.

136. T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu. A stage-wise refinement model for detecting salient objects in images. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4019–4028, 2017.

137. T. Wang, Y. Piao, X. Li, L. Zhang, and H. Lu. Deep learning for light field saliency detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 8838–8848, 2019.

138. W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang. Saliency object detection in the deep learning era: An in-depth survey. *arXiv preprint arXiv:1904.09146*, 2019.

139. W. Wang and J. Shen. Deep visual attention prediction. *IEEE Transactions on Image Processing*, 27(5):2368–2378, 2017.

140. W. Wang, J. Shen, and L. Shao. Video salient object detection via fully convolutional networks. *IEEE Transactions on Image Processing*, 27(1):38–49, 2017.

141. W. Wang, J. Shen, R. Yang, and F. Porikli. Saliency-aware video object segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(1):20–33, 2017.

142. W. Wang, S. Zhao, J. Shen, S. C. Hoi, and A. Borji. Saliency object detection with pyramid attention and salient edges. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1448–1457, 2019.

143. X. Wang, Y. Dong, Q. Zhang, and Q. Wang. Region-based depth feature descriptor for saliency detection on light field. *Multimedia Tools and Applications*, 2020.

144. Y. Wang, Y. Li, J. H. Elder, H. Lu, and R. Wu. Synergistic saliency and depth prediction for RGB-D saliency detection. *arXiv preprint arXiv:2007.01711*, 2020.

145. Y.-H. Wu, S.-H. Gao, J. Mei, J. Xu, D.-P. Fan, C.-W. Zhao, and M.-M. Cheng. Jcs: An explainable covid-19 diagnosis system by joint classification and segmentation. *arXiv preprint arXiv:2004.07054*, 2020.

146. F. Xiao, B. Li, Y. Peng, C. Cao, K. Hu, and X. Gao. Multi-modal weights sharing and hierarchical feature fusion for rgbd salient object detection. *IEEE Access*, 8:26602–26611, 2020.

147. C. Xu, D. Tao, and C. Xu. Multi-view learning with incomplete views. *IEEE Transactions on Image Processing*, 24(12):5812–5825, 2015.

148. H. Xue, Y. Gu, Y. Li, and J. Yang. Rgb-d saliency detection via mutual guided manifold ranking. In *Proceedings of IEEE International Conference on Image Processing*, pages 666–670. IEEE, 2015.

149. P. Yan, G. Li, Y. Xie, Z. Li, C. Wang, T. Chen, and L. Lin. Semi-supervised video salient object detection using pseudo-labels. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7284–7293, 2019.

150. Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1162, 2013.

151. C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019.

152. Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian. Deep modular co-attention networks for visual question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6281–6290, 2019.

153. Y. Zeng, Y. Zhuge, H. Lu, and L. Zhang. Joint learning of saliency detection and weakly supervised semantic segmentation. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 7223–7233. Springer, 2019.

154. Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, and Y. Yu. Multi-source weak supervision for saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6074–6083, 2019.

155. D. Zhang, H. Fu, J. Han, A. Borji, and X. Li. A review of co-saliency detection algorithms: Fundamentals, applications, and challenges. *ACM Transactions on Intelligent Systems and Technology*, 9(4):1–31, 2018.

156. D. Zhang, J. Han, and Y. Zhang. Supervision by fusion: Towards unsupervised learning of deep salient object detector. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4048–4056, 2017.

157. D. Zhang, D. Meng, and J. Han. Co-saliency detection via a self-paced multiple-instance learning framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):865–878, 2016.

158. D. Zhang, D. Meng, L. Zhao, and J. Han. Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. *arXiv preprint arXiv:1703.01290*, 2017.

159. H. Zhang, J. Lei, X. Fan, M. Wu, P. Zhang, and S. Bu. Depth combined saliency detection based on
region contrast model. In Proceedings of International Conference on Computer Science & Education, pages 763–766. IEEE, 2012.

[160] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. S. Saleh, T. Zhang, and N. Barnes. Uc-net: uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.

[161] J. Zhang, W. Li, P. Wang, P. Ogunbona, S. Liu, and C. Tang. A large scale rgb-d dataset for action recognition. In Proceedings of the International Workshop on Understanding Human Activities through 3D Sensors, pages 101–114. Springer, 2016.

[162] J. Zhang, Y. Liu, S. Zhang, R. Poppe, and M. Wang. Light field saliency detection with deep convolutional networks. IEEE Transactions on Image Processing, 29:4421–4434, 2020.

[163] J. Zhang, M. Wang, J. Gao, Y. Wang, X. Zhang, and X. Wu. Saliency detection with a deeper investigation of light field. In Proceedings of the International Joint Conference on Artificial Intelligence, pages 2212–2218, 2015.

[164] J. Zhang, M. Wang, L. Lin, X. Yang, J. Gao, and Y. Rui. Saliency detection on light field: A multi-cue approach. ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM), 13(3):1–22, 2017.

[165] M. Zhang, W. Ji, Y. Piao, J. Li, Y. Zhang, S. Xu, and H. Lu. Lfnet: Light field fusion network for salient object detection. IEEE Transactions on Image Processing, 29:6276–6287, 2020.

[166] M. Zhang, J. Li, J. WEI, Y. Piao, and H. Lu. Memory-oriented decoder for light field salient object detection. In Proceedings of the International Conference on Neural Information Processing Systems, pages 896–906, 2019.

[167] M. Zhang, W. Ren, Y. Piao, Z. Rong, and H. Lu. Select, supplement and focus for RGB-D saliency detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2020.

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

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

[170] Q. Zhang, N. Huang, L. Yao, D. Zhang, C. Shan, and J. Han. Rgb-t salient object detection via fusing multi-level cnn features. IEEE Transactions on Image Processing, 29:3321–3335, 2019.

[171] Z. Zhang, Z. Lin, J. Xu, W. Jin, S.-P. Lu, and D.-P. Fan. Bilateral attention network for rgb-d salient object detection. arXiv preprint arXiv:2004.14582, 2020.

[172] J. Zhao, Y. Zhao, J. Li, and X. Chen. Is depth really necessary for salient object detection. arXiv preprint arXiv:2006.00269, 2020.

[173] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, and L. Zhang. Contrast prior and fluid pyramid integration for RGBD salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3927–3936, 2019.

[174] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng. Egnet: Edge guidance network for salient object detection. In Proceedings of the IEEE International Conference on Computer Vision, pages 8779–8788, 2019.

[175] R. Zhao, W. Oyang, and X. Wang. Person re-identification by saliency learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 39(2):356–370, 2016.

[176] T. Zhao and X. Wu. Pyramid feature attention network for saliency detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 3085–3094, 2019.

[177] X. Zhao, L. Zhang, Y. Pang, H. Lu, and L. Zhang. A single stream network for robust and real-time rgb-d salient object detection. In Proceedings of the European Conference on Computer Vision. Springer, 2020.

[178] Z.-Q. Zhao, P. Zheng, S.-t. Xu, and X. Wu. Object detection with deep learning: A review. IEEE Transactions on Neural Networks and Learning Systems, 30(11):3212–3232, 2019.

[179] L. Zhou, Z. Yang, Q. Yuan, Z. Zhou, and D. Hu. Saliency region detection via integrating diffusion-based compactness and local contrast. IEEE Transactions on Image Processing, 24(11):3308–3320, 2015.

[180] T. Zhou, H. Fu, G. Chen, J. Shen, and L. Shao. Hinet: hybrid-fusion network for multi-modal MR image synthesis. IEEE Transactions on Medical Imaging, 2020.

[181] T. Zhou, M. Liu, K.-H. Thung, and D. Shen. Latent representation learning for alzheimer’s disease diagnosis with incomplete multi-modality neuroimaging and genetic data. IEEE Transactions on Medical Imaging, 38(10):2411–2422, 2019.

[182] T. Zhou, K.-H. Thung, M. Liu, F. Shi, C. Zhang, and D. Shen. Multi-modal latent space inducing ensemble svm classifier for early dementia diagnosis with neuroimaging data. Medical Image Analysis, 60:101630, 2020.

[183] T. Zhou, K.-H. Thung, X. Zhu, and D. Shen. Effective feature learning and fusion of multimodality data using stage-wise deep neural network for dementia diagnosis. Human Brain Mapping, 40(3):1001–1016.
[184] W. Zhou, Y. Chen, C. Liu, and L. Yu. GFNet: Gate fusion network with res2net for detecting salient objects in rgb-d images. *IEEE Signal Processing Letters*, 2020.

[185] W. Zhou, Y. Lv, J. Lei, and L. Yu. Global and local-contrast guides content-aware fusion for rgb-d saliency prediction. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 2019.

[186] X. Zhou, G. Li, C. Gong, Z. Liu, and J. Zhang. Attention-guided RGBD saliency detection using appearance information. *Image and Vision Computing*, 95:103888, 2020.

[187] Y. Zhou, S. Huo, W. Xiang, C. Hou, and S.-Y. Kung. Semi-supervised salient object detection using a linear feedback control system model. *IEEE Transactions on Cybernetics*, 49(4):1173–1185, 2018.

[188] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li. PDNet: Prior-model guided depth-enhanced network for salient object detection. In *Proceedings of the IEEE International Conference on Multimedia and Expo*, pages 199–204. IEEE, 2019.

[189] C. Zhu and G. Li. A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 3008–3014, 2017.

[190] C. Zhu, G. Li, W. Wang, and R. Wang. An innovative salient object detection using center-dark channel prior. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1509–1515, 2017.

[191] D. Zhu, L. Dai, Y. Luo, G. Zhang, X. Shao, L. Itti, and J. Lu. Multi-scale adversarial feature learning for saliency detection. *Symmetry*, 10(10):457, 2018.

[192] J.-Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu. Unsupervised object class discovery via saliency-guided multiple class learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(4):862–875, 2014.

[193] L. Zhu, Z. Cao, Z. Fang, Y. Xiao, J. Wu, H. Deng, and J. Liu. Selective features for rgb-d saliency. In *Proceedings of Chinese Automation Congress*, pages 512–517. IEEE, 2015.

**Tao Zhou** received Ph.D. degree in Pattern Recognition and Intelligent System from the Institute of Image Processing and Pattern Recognition, Shanghai Jiao Tong University, in 2016. He is currently a Research Scientist at the Inception Institute of Artificial Intelligence, Abu Dhabi, United Arab Emirates. His research interests include machine learning, computer vision, and medical image analysis.

**Deng-Ping Fan** received his Ph.D. degree from the Nankai University in 2019. He joined the Inception Institute of Artificial Intelligence (IIAI) in 2019. He has published about 20 top journal and conference papers such as CVPR, ICCV, etc. His research interests include computer vision, deep learning, and saliency detection, especially on co-salient object detection, RGB salient object detection, RGB-D salient object detection, and video salient object detection.

**Ming-Ming Cheng** (Senior Member, IEEE) received his Ph.D. degree from Tsinghua University in 2012. He then did 2 years research fellow, with Prof. Philip Torr in Oxford. He is now a professor at Nankai University, leading the Media Computing Lab. His research interests include computer graphics, machine learning, computer vision, and image processing. He is an Associate Editor of IEEE TIP. He received several research awards, including the ACM China Rising Star Award, the IBM Global SUR Award, etc.

**Jianbing Shen** is currently acting as the Lead Scientist with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates. He is also a Full Professor with the School of Computer Science, Beijing Institute of Technology. He has published about 100 journal and conference papers such as IEEE TPAMI, CVPR, and ICCV. His research interests include computer vision and deep learning. He is an Associate Editor of IEEE TNNLS, IEEE TIP, etc.

**Ling Shao** is the CEO and Chief Scientist of the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates. His research interests include computer vision, machine learning and medical imaging. He is an associate editor of IEEE TRANSACTIONS ON IMAGE PROCESSING, IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, and several other journals. He is a fellow of the International Association of Pattern Recognition, the Institution of Engineering and Technology and the British Computer Society.