Rethinking RGB-D Salient Object Detection: Models, Data Sets, and Large-Scale Benchmarks

Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng

Abstract—The use of RGB-D information for salient object detection has been extensively explored in recent years. However, relatively few efforts have been put towards modelling salient object detection in real-world human activity scenes with RGB-D. In this work, we fill the gap by making the following contributions to RGB-D salient object detection. (1) We carefully collect a new SIP (salient person) dataset, which consists of ∼1K high-resolution images that cover diverse real-world scenes from various viewpoints, poses, occlusions, illuminations, and backgrounds. (2) We conduct a large-scale (and, so far, the most comprehensive) benchmark comparing contemporary methods, which has long been missing in the field and can serve as a baseline for future research. We systematically summarize 32 popular models, and evaluate 18 parts of 32 models on seven datasets containing a total of about 97K images. (3) We propose a simple general architecture, called Deep Depth-Depurator Network (D3Net). It consists of a depth depurator unit (DDU) and a three-stream feature learning module (FLM), which performs low-quality depth map filtering and cross-modal feature learning respectively. These components form a nested structure and are elaborately designed to be learned jointly. D3Net exceeds the performance of any prior contenders across all five metrics under consideration, thus serving as a strong model to advance research in this field. We also demonstrate that D3Net can be used to efficiently extract salient object masks from real scenes, enabling effective background changing application with a speed of 65fps on a single GPU. All the saliency maps, our new SIP dataset, the D3Net model, and the evaluation tools are publicly available at https://github.com/DengPingFan/D3NetBenchmark.

Index Terms—Benchmark, SIP Dataset, Salient Object Detection, Saliency, RGB-D.

I. INTRODUCTION

HOW to take high-quality photos has become one of the most important competition points among mobile phone manufacturers. Salient object detection (SOD) methods [1]–[18] have been incorporated into mobile phones and been widely used for creating perfect portraits by automatically adding large aperture and other enhancement effects. While existing SOD methods [19]–[35] have achieved remarkable success, most of them only rely on RGB images and ignore the important depth information, which is widely available in modern smartphones (e.g., iPhone X, Huawei Mate10, and Samsung Galaxy S10). Thus, fully utilizing RGB-D information for SOD detection has recently attracted significant research attention [36]–[51].

One of the primary goals of existing smartphone cameras is to identify humans in visual scenes, through either coarse, bounding-box-level, or instance-level; segmentation. To this end, intelligence solutions, such as RGB-D saliency detecting techniques have gained considerable attention.

However, most existing RGB-D based SOD methods are tested on RGB-D images taken by Kinect [52] or a light field camera [53], or estimated by optical flow [54], which have different characteristics from actual smartphone cameras. Since humans are the key subjects of photographs taken with smartphones, a human-oriented RGB-D dataset featuring realistic, in-the-wild images would be more useful for mobile manufacturers. Despite the effort of some authors [37], [39] to augment their scenes with additional objects, a human-centered RGB-D dataset for salient object detection does not yet exist.

Furthermore, although depth maps provide important complementary information for identifying salient objects, the low-quality versions often cause wrong detections [55]. While existing RGB-D based SOD models typically fuse RGB and depth features by different strategies [51]. There is no model that explicitly/automatically discard the low-quality depth map in the RGB-D SOD field. We believe such models have a high potential for driving this field forward.

In addition to the limitations of current RGB-D datasets and models already mentioned, most RGB-D studies also suffer from several other common constraints, including:

Sufficiency. Only a limited number of datasets (1~4) have been benchmarked in recent papers [39], [56] (Table II). The
Figure 2. Representative subsets in our SIP. The images in SIP are grouped into eight subsets according to background objects (i.e., grass, car, barrier, road, sign, tree, flower, and other), different lighting conditions (i.e., low-light, sunny with clear object boundary) and various number of objects (i.e., 1, 2, ≥3).

Figure 3. Examples of images, depth maps and annotations (i.e., object-level, instance-level) in our SIP dataset with different numbers of salient objects, object sizes, object positions, scene complexities, and lighting conditions. Note that the “RGB” and “Gray” images are captured by two different monocular cameras from short distances. Thus, the “Gray” images are slightly different from the grayscale images obtained from colorful (RGB) image. Our SIP dataset provides a new direction such as depth estimating from “RGB” and “Gray” images, and instance-level RGB-D salient object detection.

generalizability of models cannot be properly accessed with such a small number of datasets.

Completeness. F-measure [57], MAE, and PR (precision & recall) Curve are the three most widely-used metrics in existing works. However, as suggested by [58], [59], these metrics essentially act at a pixel-level. It is thus difficult to draw thorough and reliable conclusions from quantitative evaluations [60].

Fairness. Some works [49], [51], [61] use the same F-measure metric, but do not explicitly describe which statistic (e.g., mean or max) was used, easily resulting in unfair comparison and inconsistent performance. Meanwhile, the different threshold strategies for F-measure (e.g., 255 varied thresholds [51], [61], [62], adaptive saliency threshold [39], [41], and self-adaptive threshold [43]) will result in different performance. It is thus of crucial need to provide a fair comparison of RGB-D based SOD models by extensively evaluating them with same metrics on a standard leaderboard.

A. Contribution

To address the above-mentioned problems, we provide three distinct contributions.

1) We have built a new Salient Person (SIP) dataset (see Fig. 2, Fig. 3). It consists of 929 accurately annotated high-resolution images which are designed to contain multiple salient persons per image. It is worth mentioning that the depth maps are captured by a real smartphone. We believe such a dataset
is highly valuable and will facilitate the application of RGB-D models to mobile devices. Besides, the dataset is carefully designed to cover diverse scenes, various challenging situations (e.g., occlusion, appearance change), and elaborately annotated with pixel-level ground truths (GT). Another discriminative feature of our SIP dataset is the availability of both RGB and grayscale images captured by a binocular camera, which can benefit a broad number of research directions, such as, stereo matching, depth estimation, human-centered detection, etc.

(2) With the proposed SIP and six existing RGB-D datasets [37]–[39], [63], [64], [66], we provide a more comprehensive comparison of 32 classical RGB-D salient object detection models and present the large-scale (~97K images) fair evaluation of 18 state-of-the-art (SOTA) algorithms [37]–[47], [49], [55], [67]–[69], making our study a good all-around RGB-D benchmark. To further promote the development of this field, we additionally provide an online evaluation platform with the preserved test set.

(3) We propose a simple general model called Deep Depth-Depurator Network (D³Net), which learns to automatically discard low-quality depth maps using a novel depth depurator unit (DDU). Thanks to the gate connection mechanism, our D³Net can predict salient objects accurately. Extensive experiments demonstrate that our D³Net remarkably outperforms prior work on many challenging datasets. Such a general framework design helps to learn cross-modality features from RGB images and depth maps.

Our contributions offer a systematic benchmark equipped with the basic tools for comprehensive assessment of RGB-D models, offering deep insight into the task of RGB-D based modelling and encouraging future research in this direction.

B. Organization

In § II, we first review current datasets for RGB-D salient object detection, as well as representative models for this task. Then, we present details on the proposed salient person dataset SIP in § III. In § IV, we describe our D³Net model for RGB-D salient object detection by explicitly filtering out the low-quality depth maps.

In § V, we provide both a quantitative and qualitative experimental analysis of the proposed algorithm. Specifically, in § V-A, we offer more details on our experimental settings, including the benchmarked models, datasets and runtime. In § V-B, five evaluation metrics (E-measure [59], S-measure [58], MAE, PR Curve, and F-measure [57]) are described in detail. In § V-C, we provide the mean statistics over different datasets and summarize them in Table IV. Comparison results of 18 SOTA RGB-D based SOD models over seven datasets, namely STERE [63], LFSD [64], DES [38], NLPR [39], NJU2K [37], SSD [66], and SIP (Ours) clearly demonstrate the robustness and efficiency of our D³Net model.

In § V-D, we provide a performance comparison between traditional and deep models. We also discuss the experimental results in more detail. In § V-E, we provide visualizations of the results and present saliency maps generated for various challenging scenes. In § VI, we discuss some potential applications about human activities and provide an interesting and realistic use scenario of D³Net in a background changing application. To better understand the contributions of DDU in the proposed D³Net, in § VII, we present the upper and lower bound of the DDU. All in all, the extensive experimental results clearly demonstrate that our D³Net model exceeds the performance of any prior competitors across five different metrics. In § VII-B, we discuss the limitations of this work. Finally, § VIII concludes the paper.

II. RELATED WORKS

A. RGB-D Datasets

Over the past few years, several RGB-D datasets have been constructed for SOD. Some statistics of these datasets are shown in Table I. Specifically, the STERE [63] dataset was the first collection of stereoscopic photos in this field. GIT [36], LFSD [64] and DES [64] are three small-sized datasets. GIT and LFSD were designed with specific purposes in mind, e.g., saliency-based segmentation of generic objects, and saliency detection on the light field. DES has 135 indoor images captured by Microsoft Kinect [52]. Although these datasets have advanced the field to various degrees, they are severely restricted by their small scale or low resolution. To overcome these barriers, Peng et al. created NLPR [39], a large-scale RGB-D dataset with a resolution of 640×480. Later, Ju et al. built NJU2K [37], which has become one of the most popular RGB-D datasets. The recent SSD [66] dataset partially remedied the resolution restriction of NLPR and NJU2K. However, it only contains 80 images. Despite the progress made by existing...
RGB-D datasets, they still suffer from the common limitation of not capturing depth maps in the real smartphones, making them unsuitable for reflecting real environmental conditions (e.g., lighting or distance to object).

Compared to previous datasets, the proposed SIP dataset has three fundamental differences:

- It includes 929 images with many challenging situations [83] (e.g., dark background, occlusion, appearance change, and out-of-view) from various outdoor scenarios.
- The RGB, grayscale images, and estimated depth maps are captured by a smartphone with a dual-camera. Due to the predominant application of SOD to human subjects on mobile phones, we also focus on this and thus and, for the first time, emphasize the salient persons in the real-world scenes.
- A detailed quantitative analysis is presented for the quality of the dataset (e.g., center bias, object size distribution, etc.), which was not carefully investigated in previous RGB-D based studies.

B. RGB-D Models

Traditional models rely heavily on hand-crafted features (e.g., contrast [38], [39], [73], [75], shape [36]). By embedding the classical principles (e.g., spatial bias [38], center-dark channel [46], 3D [77], background [40], [47]), difference of Gaussian [37], region classification [62], SVM [45], [73], graph knowledge [55], cellular automata [42], and Markov random field [40], [75], these models show that specific hand-crafted features can lead to decent performance. Several studies have also explored methods of integrating RGB and depth features via various combination strategies, using, for instance, angular densities [41], random forest regressors [45], [62], and minimum barrier distances [77]. More details are shown in Table II.

To overcome the limited expression ability of hand-crafted features, recent works [43], [44], [48], [49], [51], [61], [76], [78], [80]–[82] have proposed to introduce CNNs to infer salient objects from RGB-D data. BED [76] and DF [44] are two pioneering works for this, which introduced deep learning technology into the RGB-D based SOD task. More recently, Huang et al. developed a more efficient end-to-end model [78] with a modified loss function. To address the shortage of training data, Zhu et al. [48] presented a robust prior model with a guided depth-enhancement module for SOD. In addition, Chen et al. developed a series of novel approaches for this field, such as hidden structure transfer [43], a complementarity fusion module [49], an attention-aware component [80], [82], and dilated convolutions [81]. Nevertheless, these works, to the best of our knowledge, are dedicated to extracting general depth features/information.

We argue that not all information in a depth map is informative for SOD, and low-quality depth maps often introduce significant noise (1st row in Fig. 1). Thus, we instead design a simple general framework D²Net, which is equipped with a depth-depurator unit to explicitly exclude low-quality depth maps when learning complementary feature.

III. PROPOSED DATASET

A. Dataset Overview

We introduce SIP, the first human activities oriented salient person detection dataset. Our dataset contains 929 RGB-D images belonging to eight different background scenes, under two different object boundary conditions, which portray multiple actors. Each of them wears different clothes in different images. Following [83], the images are carefully selected to cover diverse challenging cases (e.g., appearance change, occlusion, and shape complexity). Examples can be found in Fig. 2 and Fig. 3. The overall dataset can be downloaded from our website http://dpfan.net/SIPDataset/.

B. Sensors and Data Acquisition

Image Collection: We used a Huawei Mate 10 to collect our images. The Mate 10’s rear camera feature high-grade Leica SUMMILUX-H lenses with bright f/1.6 apertures and combine 12MP RGB and 20MP Monochrome (grayscale) sensors. The depth map is automatically estimated by the Mate10. We asked nine people, all dressed in different colors, to perform specific actions in real-world daily scenes. Instructions on how to perform the action to cover different challenging situations (e.g., occlusion, out-of-view) were given, but no instructions on style, angle, or speed were provided, in order to record realistic data.

Data Annotation: After capturing 5,269 images and the corresponding depth maps, we first manually selected about 2,500 images, each of which included one or multiple salient people. Following many famous SOD datasets [19], [57], [70], [71], [84]–[90], six viewers were further instructed to draw the bounding boxes (bboxes) around the most attention-grabbing person, according to their first instinct. We adopted the voting scheme described in [39] to discard images with low voting consistency and chose top 1,000 most satisfactory images. Another five annotators were then introduced to label accurate silhouettes of the salient objects according to the bboxes. We discard some images with low-quality annotations and finally obtained the 929 images with high-quality ground-truth annotations.

C. Dataset Statistics

Center Bias: Center bias has been identified as one of the most significant biases of saliency detection datasets [91]. It occurs because subjects tend to look at the center of a screen [92]. As noted in [83], simply overlapping all of the maps in the dataset cannot well describe the degree of center bias.

Following [83], we present the statistics of two distance $R_o$ and $R_m$ in Fig. 4 (a & b), where $R_o$ and $R_m$ indicate how far an object center and margin (farthest) point in an object are from the image center, respectively. The center biases of our SIP and existing [36]–[39], [63], [64], [66] datasets are shown in Fig. 4 (a & b). Except for our SIP and two small-scale datasets (GIT and SSD), most datasets present a high degree of center bias, i.e. the center of the object is close to the image center.

Size of Objects: We define object size as the ratio of salient object pixels to the total number of pixels in the image. The
Table II
Comparision of 31 classical RGB-D based SOD algorithms and the proposed baseline (D^Net). Train/Val Set. (#) = Training or Validation Set: NLR = NLPR [39], NJU = NJU2K [37], MK = MSA10K [70], O = MK + DUS [71]. Basic: 4Priors = 4 priors, e.g., Region, Background, Depth, and Surface Orientation Prior. IPT = Initialization Parameters Transfer. LGBS Priors = Local Contrast, Global Context, Background, and Spatial Prior. RFR [72] = Random Forest Regressor. MCFM = Multi-constraint Feature Matching. CLP = Cross Label Propagation. Type: T = Traditional, D = Deep Learning. SP = SuperPixel. Whether or not use the superpixel algorithm. E-Measure: The range of scores over the seven datasets in Table IV. Evaluation tools: https://github.com/DengPingFan/E-measure.

| No. | Model | Year | Pub. | Train/Val Set. (#) | Test (#) | Basic | Type | SP | E-measure | [%] |
|-----|-------|------|------|-------------------|----------|-------|------|----|-----------|-----|
| 1   | LS    | [36] | 2013 | BMVC              | Without training dataset | One    | Markov Random Field | T  | ✓          | Not Available |
| 2   | RC    | [33] | 2013 | BMVC              | Without training dataset | One    | Region Contrast, SVM [74] | T  | ✓          | Not Available |
| 3   | LHM   | [39] | 2014 | ECCV              | Without training dataset | One    | Multi-Context Contrast | T  | ✓          | 0.653–0.771 |
| 4   | DESM  | [38] | 2014 | ICIMCS            | Without training dataset | One    | Color/Depth Contrast, Spatial Bias Prior | T  | 0.770–0.868 |
| 5   | ACSD  | [37] | 2014 | ICP               | Without training dataset | One    | Difference of Gaussian | T  | 0.780–0.850 |
| 6   | SRRD  | [35] | 2014 | DSP               | Without training dataset | One    | Weighted Color Contrast | T  | Not available |
| 7   | GP    | [60] | 2015 | CVPRW             | Without training dataset | Two    | Two Markov Random Field, 4Priors | T  | 0.670–0.824 |
| 8   | PRC   | [62] | 2016 | Access            | Without training dataset | Two    | Region Classification, RFR | T  | Not available |
| 9   | LBE   | [51] | 2016 | CVPR              | Without training dataset | Two    | Two Angular Density Component | T  | 0.736–0.890 |
| 10  | DMCE  | [55] | 2016 | SPL               | Without training dataset | Two    | Two Depth Confidence, Compactness, Graph | T  | 0.743–0.856 |
| 11  | SE    | [42] | 2016 | ICME              | Without training dataset | Two    | Two Cellular Automata | T  | 0.771–0.856 |
| 12  | MCLP  | [67] | 2017 | Cybernetic        | Without training dataset | Two    | Two Addition, Deletion and Iteration Scheme | T  | ✓          | Not Available |
| 13  | TPF   | [66] | 2017 | ICCVV             | Without training dataset | Four   | Cellular Automata, Optical Flow | T  | ✓          | Not Available |
| 14  | DCOP  | [65] | 2017 | ICCVV             | Without training dataset | Two    | Two Center-dark Channel Prior | T  | 0.700–0.820 |
| 15  | DF    | [54] | 2017 | TIP               | NLR (0.5Kj) + NJU (1.0K) | Three  | Laplacian Propagation, LGBS Priors | D  | 0.759–0.880 |
| 16  | BED   | [76] | 2017 | ICCVV             | NLR (0.00Kj) + NJU (1.6K) + MK (9K) | Two    | Two Background Enclosure Distribution | D  | Not available |
| 17  | MDSP  | [35] | 2017 | TIP               | NLR (0.5Kj) + NJU (0.5K) | Two    | Two SVM [74], RFR, Ultrametric Contour Map | T  | 0.779–0.885 |
| 18  | MIP   | [77] | 2017 | SPL               | Without training dataset | One    | One Minimum Barrier Distance, 3D prior | T  | Not available |
| 19  | Review| [56] | 2018 | TCSVT             | Without training dataset | Two    | Two Without model introduced | T  | Not available |
| 20  | HSCS  | [68] | 2018 | TMM               | Without training dataset | Two    | Two Hierarchical Sparsity, Energy Function | T  | ✓          | Not Available |
| 21  | ICS   | [69] | 2018 | TIP               | Without training dataset | One    | One MCFM, CLP | T  | ✓          | Not available |
| 22  | CTD   | [47] | 2018 | NC                | Without training dataset | One    | One Background Prior | T  | 0.698–0.830 |
| 23  | SCDL  | [78] | 2018 | DSP               | NLR (0.5Kj) + NJU (1.0K) | Two    | Two Silhouette Feature, Spatial Coherence Loss | D  | Not available |
| 24  | PCF   | [69] | 2018 | CVPR              | NLR (0.5Kj) + NJU (1.5K) | Three  | Three Complementarity-Aware Fusion module [49] | D  | 0.827–0.925 |
| 25  | CTMF  | [53] | 2018 | Cybernetic        | NLR (0.5Kj) + NJU (1.4K) | Four   | HHA [75], IPT, Hidden Structure Transfer | D  | 0.829–0.932 |
| 26  | ACCP  | [50] | 2018 | CBNet             | NLR (0.5Kj) + NJU (1.4K) | Three  | Three Attention-Aware | D  | Not available |
| 27  | PNMT  | [38] | 2018 | ICMIE             | NLR (0.5Kj) + NJU (1.5K) + O (21K) | Five   | Five Depth-Enhanced Net [48] | D  | Not available |
| 28  | AFNet | [61] | 2019 | Access            | NLR (0.70Kj) + NJU (1.5K) | Three  | Three Switch map, Edge-Aware loss | D  | 0.807–0.887 |
| 29  | MCMC  | [51] | 2019 | NJU2K             | NLR (0.70Kj) + NJU (1.5K) | Three  | Three HHA [75], Dilated Convolutional | D  | 0.833–0.928 |
| 30  | TANET | [52] | 2019 | TIP               | NLR (0.70Kj) + NJU (1.5K) | Three  | Three Attention-Aware Multi-Modal Fusion | D  | 0.847–0.941 |
| 31  | CPFP  | [51] | 2019 | CVPR              | NLR (0.70Kj) + NJU (1.5K) | Five   | Five Contrast Prior, Fluid Pyramid | D  | 0.852–0.932 |

| 32  | D^Net (Ours) | 2020 | NLR (0.70Kj) + NJU (1.5K) | Seven Depth Depurator Unit | D | 0.862–0.953 |

Figure 4. (a) Distribution of normalized object center distance from image center. (b) Distribution of normalized object margin (farthest point in an object) distance from image center. (c) Distribution of normalized object size.

Table III
Statistics regarding camera/object motions and salient object instance numbers in SIP dataset.

| SIP (Ours) | # Img | Background Objects | Object Boundary | # Object |
|------------|-------|--------------------|-----------------|---------|
|            |       | car | flower | grass | road | tree | signs | barrier | other | dark | clear | ≥3 |
|            | 107   | 9  | 154  | 140  | 97   | 25   | 366   | 32      |       | 162  | 787  |    |
|            | 591   | 159 | 179  |       |      |      |       |         |       |      |      |    |

distribution (Fig. 4 (c)) of normalized object size in SIP are 0.48%~66.85% (avg.: 20.43%).

Background Objects: As summarized in Table III, SIP includes diverse background objects (e.g., cars, trees, and grass). Models tested on such a dataset would likely be able to handle realistic scenes better and thus be more practical.

Object boundary Conditions: In Table III, we show different object boundary conditions (e.g., dark and clear) in
our SIP dataset. One example of a dark condition, which often occurs in daily scenes, can be found in Fig. 3. The depth maps obtained in low-light conditions inevitably introduce more challenges for detecting salient objects.

Number of Salient Object: From Table I, we note that existing datasets fall short in their numbers of salient objects (e.g., they often only have one). Previous studies [93], however, have shown that humans can accurately enumerate up to at least five objects without counting. Thus, our SIP is designed to contain up to five salient objects per-image. The statistics of labelled objects in each image are shown in Table III (# Object).

IV. PROPOSED MODEL

According to motivation described in Fig. 1, cross-modality feature extraction and depth filter unit are highly desired; therefore we proposed the simple general D³Net model (illustrated in Fig. 5) which contains two components, e.g., a three-stream feature learning module (§ IV-A) and a depth depurator unit (§ IV-B). The FLM (feature learning module) is utilized to extract the features from different modality. While the DDU (depth depurator unit) is acting as a gate to explicitly filter out the low-quality depth maps. If DDU decides to filter out this depth map, the data flow will pass along with the RgbNet. These components form a nested structure and are elaborately designed (e.g., gate connection in DDU) to automatically learn the salient object from the RGB image and Depth image jointly.

A. Feature Learning Module

Most existing models [94]–[96] have shown significant improvement for object detectors in several applications. These models typically share a common structure of Feature Pyramid Networks (FPN) [97]. Based on this motivation, we decide to introduce this component like FPN in our D³Net baseline to efficiently extract the features in a pyramid manner. The entire D³Net model is divided into the training phase and test phase due to the DDU has opted to use only in test phase.

As shown in Fig. 5, the designed FLM appears in training and test phases. The FLM consists of three sub-networks, i.e., RgbNet, RgbdNet, and DepthNet. Note that the three sub-networks have the same structure while fed with different input channel. Specifically, each sub-network receives a re-scaled image \( I \in \{I_{rgb}, I_{rgbd}, I_{depth}\} \) with \( 224 \times 224 \) resolution. The goal of FLM is to obtain the corresponding predicted map \( S \in \{S_{rgb}, S_{rgbd}, S_{depth}\} \).

As in [97], we also use bottom-up, top-down pathway, and lateral connections to extract the features. Then the outputs will be proportionally organized at multiple levels. The FPN is independent of the backbone, thus for simplicity, we adopt the VGG-16 [98] architecture as our basic convolutional network to extract spatial features, while utilizing more powerful backbone [99] feature extractor could be explored in future. Some studies like [100] have shown that deeper layers retain more semantic information for locating objects. Based on this observation, we introduce a layer containing two 3×3 convolution kernels on the basis of the 5 layers VGG-16 structure to achieve this goal.

As shown in Fig. 6, our top-down features are built. For a specific layer (e.g., coarser layer), we first conduct a \( 2 \times \) upsampling using nearest neighbor operation. Then, the upsampled feature is concatenated with the finer feature map to obtain rich features. Before concatenated with coarse map, the finer map undergoes a \( 1 \times 1 \) Convl operation to reduce the channel. For example, let \( I_{rgbd} \in \mathbb{R}^{W \times H \times 4} \) denotes the four-dimensional feature tensor of the input of RgbdNet. Then we define a set of anchors on different layers so that we can obtain a set of pyramid feature tensors with \( C_l \times W_l \times H_l \), i.e., \( \{64 \times 224 \times 224, 128 \times 112 \times 112, 256 \times 56 \times 56, 512 \times 28 \times 28, 512 \times 14 \times 14, 32 \times 7 \times 7, 32 \times 14 \times 14, 32 \times 28 \times 28, 32 \times ... \)
which predicted map RgbdNet, respectively. S
between the a comparison unit F
in the test phase rather than in the training phase. Specially,
processing and pre-processing. We propose a simple but general
S
connections. Denote the input images with three predicted maps
shown in Fig. 5 (right), the DDU is implemented with a gate
is firstly re-sized to a fixed size (56 × 56, 32 × 112 × 112, 32 × 224 × 224) to reduce the computational complexity. As
forward a new gate connection strategy to obtain the optimal predicted map. Low-quality depth maps introduce more noise than informative cues to the prediction. The goal of gate connection is to classify depth maps into reasonable and low-quality ones and not use the poor ones in the pipeline.

As illustrated in Fig. 7 (b), a stand-alone salient object in a high-quality depth map is typically characterized by well-defined closed boundaries and shows clear double peaks in its depth distribution. The statistics of the depth maps in existing datasets [37]–[39], [63], [64], [66] also support the fact that “high quality depth maps usually contain clear objects, while the elements in low-quality depth maps are cluttered (2nd row in Fig. 7)”.

In order to reject the low-quality depth maps, we propose DDU as follows:

- More specifically, in the test phase, the RGB and depth map is firstly re-sized to a fixed size (e.g., same as the training phase 224×224) to reduce the computational complexity. As shown in Fig. 5 (right), the DDU is implemented with a gate connection. Denote the input images with three predicted maps S ∈ \{Srgb, Srgbd, Sdepth\}, then the goal of DDU is to decide which predicted map P ∈ [0,1]W×H is optimal.

\[
P = F_{ddu}(S_{rgb}, S_{rgbd}, S_{depth}).
\]  

Intuitively, there are two ways to achieve this goal, e.g., post-processing and pre-processing. We propose a simple but general post-processing scheme for DDU. The DDU is considered in the test phase rather than in the training phase. Specially, a comparison unit Fcu is leveraged to assess the similarity between the Sdepth and Srgbd generated from DepthNet and RgbdNet, respectively.

\[
F_{cu} = \begin{cases} 
1, & \delta(S_{rgbd}, S_{depth}) \leq t \\
0, & \text{otherwise} 
\end{cases}
\]  

where the δ(·) represents distance function, and t indicates a fixed threshold. Note that the comparison unit Fcu is act as an index to decide which sub-network (RgbNet or RgbdNet) should be utilized.

The key of our comparison unit is the DDU. We utilize the comparison unit Fcu as a gate connection to decide the final/optimal predicted map P. Thus, our F_{ddu} module can be formulated as:

\[
P = F_{cu} \cdot S_{rgb} + \bar{F}_{cu} \cdot S_{rgbd},
\]  

where \(\bar{F}_{cu} = 1 - F_{cu}\). The Fcu can be viewed as a fixed weight. A more elegant formulation (adaptive weight) would be a part of our future work.

**C. Implementation Details**

DDU. The key component of our D^3Net is the DDU. In this work, we show a simple yet powerful distance function formulated in (Eq. 2). We leverage the mean absolute error (MAE) metric (same as (Eq. 5)) to assess the distance between two maps. The basic motivation is that if the high-quality depth contains clear objects the DepthNet will easily detect these objects in Sdepth (see first row in Fig. 7). The higher the quality of depth map in I_{depth} the more similarity between the Srgbd and the Sdepth. In other words, the predicted map S_{rgbd} from RgbdNet have considered the feature from I_{depth}. If the quality of the depth map is low, then the predicted map from RgbdNet will quite different from the generated map from DepthNet.

We have tested a set of values of the fixed threshold t in (Eq. 2) such as, 0.01, 0.02, 0.05, 0.10, 0.15, 0.20, but have found t = 0.15 achieve the best performance.

**Loss Function.** We adopt the widely-used cross entropy loss function L to train our model:

\[
L(S, G) = -\frac{1}{N} \sum_{i=1}^{N} \left( g_i \log(s_i) + (1 - g_i) \log(1 - s_i) \right),
\]  

where \(S \in [0,1]^{224 \times 224}\) and \(G \in [0,1]^{224 \times 224}\) indicate the estimated saliency map (i.e., Srgb, Srgbd, or Sdepth) and the GT map, respectively. \(g_i \in G, s_i \in S\), and \(N\) denotes the total number of pixels.

**Training Settings.** For fair comparisons, we follow the same training settings described in [51]. We select 1485 image pairs from the NJU2K [37] and 700 image pairs from NLP [39] dataset, respectively, as the training data (Please refer to our website for the Trainlist.txt). The proposed D^3Net
is implemented using Python, with the Pytorch toolbox. We adopt Adam as the optimizer and the initial learning rate is 1e-4 and batchsize is set to 8. The total training is 30 epoch on a GTX TITAN X GPU with 12G of memory.

**Data Augmentation.** Due to the limited scale of existing datasets, we augment the training samples by flipped the images horizontally to overcome the risk of overfitting.

**V. BENCHMARKING EVALUATION RESULTS**

We benchmark about 97K images (5,398 images × 18 models) in this study, making it the largest and most comprehensive RGB-D based SOD benchmark to date.

**A. Experimental Settings**

**Models.** We benchmark 18 SOTA models (see Table IV), including 10 traditional and 8 CNN based models.

**Datasets.** We conduct our experiments on seven datasets (see Table IV). The test sets of NJU2K [37] and NLPR [39] datasets, and the whole STERE [63], DES [38], SSD [66], LFSD [68], and SIP datasets are used for testing.

**Runtime.** In Table IV, we summarize the runtime of existing approaches. The timings are tested on the same platform: Intel Xeon(R) E5-2676v3 2.4GHz×24 and GTX TITAN X. Since [43], [47], [49], [67]–[69], [80]–[82] have not released their codes, the timings are borrowed from the original papers or provided by the authors. Our D³Net does not apply post-processing (e.g., CRF), thus the computation only takes about 0.015s for a 224 × 224 image.

**B. Evaluation Metrics**

**MAE.** We follow Perazzi et al. [101] and evaluate the mean absolute error (MAE) between a real-valued saliency map \( Sal \) and a binary ground truth \( G \) for all image pixels:

\[
\text{MAE} = \frac{1}{N} \| Sal - G \|,
\]

where \( N \) is the total number of pixels. The MAE estimates the approximation degree between the saliency map and the ground truth map, and it is normalized to \([0, 1] \). The MAE provides a direct estimate of conformity between estimated and ground truth maps. However, for the MAE metric, small objects are naturally assigned smaller errors, while larger objects are given larger errors. The metric is also unable to tell where the error occurs [102].

**PR Curve.** We also follow Borji et al. [5] and provide the PR Curve. We divide a saliency map \( S \) using a fixed threshold which changes from 0 to 255. For each threshold, a pair of recall & precision scores are computed, and then combined to form a precision-recall curve that describes the model performance in different situations. The overall evaluation results for PR Curves are shown in Fig. 8 (Top) and Fig. 9 (Left).

**F-measure \( F_\beta \).** F-measure is essentially a region-based similarity metric. Following the works by Cheng and Zhang et al. [5], [103], we also provide the max F-measure using various fixed (0-255) thresholds. The overall F-measure evaluation results under different thresholds on each dataset are shown in Fig. 8 (Bottom) and Fig. 9 (Right).

**S-measure \( S_\alpha \).** Both the MAE and F-measure metrics ignore important structural information. However, behavioral vision studies have shown that the human visual system is highly sensitive to structures in scenes [58]. Thus, we additionally include the structure measure (S-measure [58]). The S-measure combines the region-aware \( (S_r) \) and object-aware \( (S_o) \) structural similarity as the final structure metric:

\[
S_\alpha = \alpha \times S_o + (1 - \alpha) \times S_r,
\]

where \( \alpha \in [0, 1] \) is the balance parameter and set to 0.5.

**E-measure \( E_\xi \).** E-measure is the recently proposed Enhanced alignment measure [59] from the binary map evaluation field. This measure is based on cognitive vision studies, and combines local pixel values with the image-level mean value in one term, jointly capturing image-level statistics and local pixel matching information. Here, we introduce max/maximal E-measure to provide a more comprehensive evaluation.

**C. Metric Statistics**

For a given metric \( \zeta \in \{ S_\alpha, F_\beta, E_\xi, M \} \) we consider different statistics. \( I_j \) denote an image from a specific dataset \( D_i \), Thus, \( D_i = \{ I_1, I_2, \ldots, I_n \} \). Let \( \zeta(I_j) \) be the metric score on image \( I_j \). The mean is the average dataset statistic defined as \( M_\zeta(D_i) = \frac{1}{|D_i|} \sum \zeta(I_j) \), where \( |D_i| \) is the total number of images on the \( D_i \) dataset. The mean statistics over different datasets are summarized in Table IV.

**D. Performance Comparison and Analysis**

**Performance of Traditional Models.** Based on the overall performances listed in Table IV, we observe that “SE [42], MDSF [45], and DCMC [55] are the top-3 traditional algorithms.” Utilizing superpixel technology, both SE and DCMC explicitly extract the region contrast features from an RGB image. In contrast, MDSF formulates SOD as a pixel-wise binary labelling problem, which is solved by SVM.

**Performance of Deep Models.** Our D³Net, CPFP [51] and TANet [82] are the top-3 deep models out of all leading methods, showing the strong feature representation ability of deep learning for this task.

**Traditional vs Deep Models.** From Table IV, we observe that most of the deep models perform better than the traditional algorithms. Interestingly, MDSF [66] outperforms two deep models (i.e., DF [44] and AFNet [61]) on the NLPR dataset.

**E. Comparison with SOTAs**

We compare our D³Net with 17 SOTA models in Table IV. In general, our model outperforms the best published result (CPFP [51]-CVPR’19) by large margins of 1.0% ~ 5.8% on six datasets. Notably, we also achieve a significant improvement of 1.4% on the proposed real-world SIP dataset.

We also report saliency maps generated on various challenging scenes to show the visual superiority of our D³Net. Some representative examples are shown in Fig. 10, such as when the structure of the salient object in the depth map is partially (e.g., the 1st, 4th, and 5th rows) or dramatically (i.e., the 2nd, 3rd rows) damaged. Specifically, in the 3rd and 5th rows, the depth
of the salient object is locally connected with background scenes. Also, the 4th row contains multiple isolated salient objects. For these challenging situations, most of the existing top competitors are unlikely to locate the salient objects due to their poor depth maps or insufficient multi-modal fusion schemes. Although CPFP [51], TANet [82], and PCF [49] can generate more correct saliency maps than others, the salient object often introduces noticeable distinct backgrounds (3rd-5th rows) or the fine details of the salient object are lost (1st row) due to the lack of a cross-modality learning ability. In contrast, our D3Net can eliminate low-quality depth maps and adaptively select complementary cues from RGB and depth images to infer the real salient object and highlight its details.

VI. APPLICATIONS

A. Human Activities

Nowadays, mobile phones generally have deep sensing cameras. With RGB-D salient object detection, users can better achieve the following functions: object extraction, a bokeh effect, mobile user recognition, etc. Many monitoring probes also have depth sensors, and RGB-D SOD can be helpful to the discovery of suspicious objects. For example, there is a lidar probe in autonomous vehicles designed to obtain depth information. RGB-D SOD is thus helpful for detecting basic objects such as pedestrians and signboards in these vehicles. There are also depth sensors in most industrial robots, so RGBD-SOD can help them better perceive the environment.
Table IV Benchmarking results of 18 leading RGB-D approaches on our SIP and FDPsix classical [37–39], [63], [64], [66] datasets. ↑ & ↓ denote larger and smaller is better, respectively. “-T” indicates the test set of the corresponding dataset. For traditional models, the statistics are based on overall datasets rather than the test set. The “Rank” denotes the ranking of each model in a specific measure. The “All Rank” indicates the overall ranking (average of each rank) in a specific dataset. The best performance is highlighted in BOLD.

| Model | 2014-2017 | 2018-2019 |
|-------|-----------|-----------|
|       | LHM | CDB | DESM | GP | CDPC | ACSD | LBE | DCMD | MDSP | SE | AFNet | CFM | MCMC | PCF | TANet | CPPF | D'Net |
| Time (s) Code | 2.130 | 1.770 | 1.298 | 3.110 | 1.200 | >0.60 | 0.718 | M&C | <0.60 | M&C | >0.570 | 10.36 | M&C | 0.030 | 0.630 | 0.050 | 0.060 | 0.070 | 0.170 | 0.015 |
| Rank | 17 | 16 | 14 | 17 | 15 | 12 | 10 | 13 | 9 | 11 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |

**DATASETS**

| Model | 2014-2017 | 2018-2019 |
|-------|-----------|-----------|
|       | LHM | CDB | DESM | GP | CDPC | ACSD | LBE | DCMD | MDSP | SE | AFNet | CFM | MCMC | PCF | TANet | CPPF | D'Net |
| Time (s) Code | 2.130 | 1.770 | 1.298 | 3.110 | 1.200 | >0.60 | 0.718 | M&C | <0.60 | M&C | >0.570 | 10.36 | M&C | 0.030 | 0.630 | 0.050 | 0.060 | 0.070 | 0.170 | 0.015 |
| Rank | 17 | 16 | 14 | 17 | 15 | 12 | 10 | 13 | 9 | 11 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
To overcome the above-mentioned drawbacks, salient object detection technology could be a potential solution. Previous similar works, such as the automatic generation of visual-textual applications [104], [105] motive us to create a background changing application for book cover layouts. We provide a prototype demo, as shown in Fig. 11. First, the user can upload an image as a candidate design image ((a) Input Image). Then, content-based image features, such as an RGB-D based saliency map, are considered in order to automatically generate salient objects. Finally, the system allows us to choose from our library of professionally designed book cover layouts ((b) Template). By combining high-level template constraints and low-level image features, we obtain the background changed book cover ((d) Results).

Since designing a complete software system is not our main focus in this article, Future researchers can follow yang et al. [104] and set our visual background image with a specified topic [105]. In stage two, the input image is resized to match the target style size and preserve the salient region according to the inference of our D$^3$Net model.

VII. DISCUSSION

Based on our comprehensive benchmarking results, we present our conclusions to the most important questions that may benefit the research community to rethink the RGB-D image for salient object detection.

A. Ablation Study.

We now provide a detailed analysis on the proposed baseline D$^3$Net model. To verify the effectiveness of the depth map filter mechanism (the DDU), we derive two ablation studies: w/o DDU and DDU, which refer to our D$^3$Net without utilizing DDU or include the DDU. For w/o DDU, we further test the performance of the three sub-network in the test phase of D$^3$Net.

In Table V, we observe that RgbdNet performs better than RgbNet on the SIP, STERE, DES, LFSD, SSD, NJU2K datasets. It indicates that the cross-modality (RGB and depth) features show strong promise for RGB-D image representation learning. In most cases, however, DepthNet has lower performance than DepthNet and RgbNet. It shows that only based on a single modality, it is difficult for the model to construct the structure of the geometry in an image.

From Table V, we also observed that the use of the DDU improves the performance (compared to RgbdNet) to a certain extent on the STERE, DES, NJU2K, and NLPR datasets. We attribute the improvement to the DDU being able to discard low-quality depth maps and select one optimal path (RgbdNet...
For the SSD dataset, however, the DDU achieves comparable performance to the single stream network (i.e., RgbdNet). It is worth mentioning that D³Net outperforms any prior approach intended for SOD, without any post-processing techniques, such as CRF, which are typically used to boost scores. In order to know the lower and upper bound of our D³Net, we additionally select the optimal path (RgbdNet or RgbNet) of the D³Net. For example, for a specific RGB ($I_{rgb}$) and depth map ($I_{depth}$), the two predicted maps i.e., $S_{rgb}$ and $S_{rgbd}$, can be assessed separately. Thus, for each input we know the best output in existing network. We aggregate all the best and worst results and achieve the upper bound and lower bound of our D³Net. From existing results listed in Table V, D³Net still has a ~1.6% performance gap on average related to the upper bound.

### B. Limitations

First, it is worth pointing out that the number of images in the SIP dataset is relatively small compared with most datasets for RGB salient object detection. Our goal behind building this dataset is to explore the potential direction of smartphone based applications. As can be seen from the benchmark results and the demo application described in § VI, salient object detection over real human activity scenes is a promising direction. We plan to keep growing the dataset with more challenging situations and various kinds of foreground persons.

Second, our simple general framework D³Net consists of three sub-networks, which may increase the memory on a lightweight device. In a real environment, several strategies can be considered to avoid this, such as replacing the backbone with MobileNet V2 [106], dimension reduction [107], or using the recently released ESPNet V2 [108] models. Third, we present the lower and upper bounds of the DDU. The optimal upper bound is obtained by feeding the input into RgbdNet or RgbNet so that the predicted map is optimal. As shown in Table V, our DDU module does not achieve the best upper bound on the current training subset. There is thus still an opportunity to design a better DDU to further improve the performance.

### VIII. CONCLUSIONS

We present systematic studies on RGB-D based salient object detection by: (1) Introducing a new human-oriented SIP dataset reflecting the realistic in-the-wild mobile use scenarios. (2) Designing a novel D³Net. (3) Conducting so far the largest-scale (~97K) benchmark. Compared with existing datasets, SIP covers several challenges (e.g., background diversity, occlusion, etc) of human in the real environments. Moreover, the proposed baseline achieves promising results. It is among the fastest methods, making it a practical solution to RGB-D salient object detection. The comprehensive benchmarking results include 32 summarized SOTAs and 18 evaluated traditional/deep models. We hope this benchmark will accelerate not only the development of this area but also others (e.g., stereo estimating/matching [109], multiple salient person detection, salient instance detection [19], sensitive object detection [110], image segmentation [111]). Note that the methods utilized in our D³Net baseline are simple and more complex components (e.g., PDC in [112]) or training strategy [113] are promising to increase the performance. In the future, we plan to incorporate recently proposed techniques e.g., the weighted triplet loss [114], hierarchical deep features [115], visual question-driven saliency [116], into our D³Net to further boost the performance. After this submission, there are many interesting models, such as UCNet [117],

### Table V

| Aspects                  | Model       | SIP (Ours) | STERE [63] | DES [38] | LFSD [64] | SSD [66] | NJU2K [37] | NLPR [39] |
|--------------------------|-------------|------------|------------|----------|-----------|----------|------------|-----------|
| w/o DDU                  | RgbdNet     | 0.831      | 0.893      | 0.881    | 0.810     | 0.839    | 0.888      | 0.911     |
|                          | RgbdNet     | 0.862      | 0.898      | 0.896    | 0.836     | 0.857    | 0.898      | 0.910     |
|                          | DepthNet    | 0.862      | 0.713      | 0.911    | 0.724     | 0.811    | 0.857      | 0.864     |
| w/ DDU                   | Lower Bound | 0.822      | 0.881      | 0.870    | 0.788     | 0.817    | 0.875      | 0.897     |
|                          | D³Net (Ours)| 0.860      | 0.899      | 0.898    | 0.825     | 0.857    | 0.900      | 0.912     |
|                          | Upper Bound | 0.872      | 0.910      | 0.907    | 0.858     | 0.879    | 0.912      | 0.924     |

Figure 12. Comparison with previous object-level datasets, which are labeled with polygons (the foot pad in NLPR [39]), coarse boundaries (i.e., the chair in DES [38]), and missed object parts (e.g., the person in SSD [66]). In contrast, the proposed object-INSTANCE-level SIP dataset is labeled with smooth, fine boundaries. More specifically, occlusions are also considered (e.g., the barrier region).
References

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

[2] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, “Detect globally, refine locally: A novel approach to saliency detection,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3172–3175.

[3] H. Fu, D. Xu, S. Lin, and J. Liu, “Object-based rgbd image co-segmentation with mutex constraint,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 4428–4436.

[4] P. Zhang, W. Liu, H. Lu, and C. Shen, “Salient object detection with lossless feature reflection and weighted structural loss,” *IEEE T. Image Process.*, 2019.

[5] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, “Salient Object Detection: A Benchmark,” *IEEE T. Image Process.*, vol. 24, no. 12, pp. 5706–5722, 2015.

[6] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, “Revisiting video saliency prediction in the deep learning era,” *IEEE T. Pattern Anal. Mach. Intell.*, 2019.

[7] D.-P. Fan, W. Wang, M.-M. Cheng, and J. Shen, “Shifting more attention to video salient object detection,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 8554–8564.

[8] Y. Zeng, Y. Zhahe, H. Lu, and L. Zhang, “Multi-source weak supervision for saliency detection,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[9] R. Wu, M. Feng, W. Guan, and D. Wang, “A mutual learning method for salient object detection with intertwined multi-supervision,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[10] L. Zhang, i. Z. Zhang, Z. Lin, H. Lu, and Y. He, “Capsal: Leveraging captioning to boost semantics for salient object detection,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[11] M. Feng, H. Lu, and E. Ding, “Attentive feedback network for boundary-aware salient object detection,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[12] X. Hu, L. Zhu, J. Qiu, C.-W. Fu, and P.-A. Heng, “Recurrently aggregating deep features for salient object detection,” in *AAAI Conference on Artificial Intelligence*, 2018.

[13] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, “Salient object detection with recurrent fully convolutional networks,” *IEEE T. Pattern Anal. Mach. Intell.*, 2018.

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

[15] N. Liu, J. Han, and M.-H. Yang, “PiCANet: Learning Pixel-wise Contextual Attention for Saliency Detection,” in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 3089–3098.

[16] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qiu, G. Han, and P.-A. Heng, “R3Net: recurrent residual refinement network for saliency detection,” in *International Joint Conference on Artificial Intelligence*, 2018, pp. 684–690.

[17] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, “Deeply supervised salient object detection with short connections,” *IEEE T. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 815–828, 2019.

[18] D. Tao, J. Cheng, M. Song, and X. Lin, “Manifold ranking-based matrix factorization for saliency detection,” *IEEE T. Neur. Net. Lear.*, vol. 27, no. 6, pp. 1122–1134, 2015.
saliency fusion and bootstrap learning,” IEEE T. Image Process., vol. 26, no. 9, pp. 4204–4216, 2017.

[46] C. Zhu, G. Li, W. Wang, and R. Wang, “An innovative salient object detection using center-dark channel prior,” in Int. Conf. Comput. Vis. Worksh., 2017.

[47] F. Liang, L. Duan, W. Ma, Y. Qiao, Z. Cai, and L. Qing, “Stereoscopic saliency model using contrast and depth-guided background priors,” Neurocomputing, vol. 275, pp. 2227–2238, 2018.

[48] C. Zhu, X. Cai, K. Huang, T. H. Li, and G. Li, “Pdnet: Prior-model guided depth-enhanced network for salient object detection,” in Int. Conf. Multimedia and Expo, 2019.

[49] H. Chen and Y. Li, “Progressively Complementarity-Aware Fusion Network for RGB-D Salient Object Detection,” in IEEE Conf. Comput. Vis. Pattern Recog., 2018, pp. 3051–3060.

[50] W. Wang, J. Shen, Y. Yu, and K.-L. Ma, “Stereoscopic thumbnail creation via efficient stereo saliency detection,” IEEE T. Image Process. Gr., vol. 23, no. 8, pp. 2014–2027, 2017.

[51] 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 IEEE Conf. Comput. Vis. Pattern Recog., 2019.

[52] Z. Zhang, “Microsoft kinect sensor and its effect,” IEEE T. Multimedia, vol. 19, no. 2, pp. 4–10, 2012.

[53] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned visual saliency detection with comprehensive information,” IEEE T. Circuit Syst. Video Technol., 2018.

[54] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, “Review of visual saliency detection with comprehensive information,” IEEE T. Circuits Syst. Video Technol., 2018.

[55] C. Liu, J. Yuen, and A. Torralba, “Sift flow: Dense correspondence across scenes and its applications,” IEEE T. Pattern Anal. Mach. Intell., vol. 33, no. 5, pp. 978–994, 2011.

[56] 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, vol. 23, no. 6, pp. 819–823, 2016.

[57] R. Cong, J. Lei, H. Fu, M.-M. Cheng, W. Lin, and Q. Huang, “Review of visual saliency detection with comprehensive information,” IEEE T. Circuits Syst. Video Technol., 2018.

[58] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” in IEEE Conf. Comput. Vis. Pattern Recog., 2009, pp. 1597–1604.

[59] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, “Structure-measure: A new way to evaluate foreground maps,” in Int. Conf. Comput. Vis., 2017, pp. 4548–4557.

[60] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, “Enhanced-alignment Measure for Binary Foreground Map Evaluation,” in International Joint Conferences on Artificial Intelligence, 2018, pp. 698–704.

[61] M. Margolin, L. Zelnik-Manor, and A. Tal, “How to evaluate foreground maps?” in IEEE Conf. Comput. Vis. Pattern Recog., 2014, pp. 248–255.

[62] N. Wang and X. Gong, “Adaptive Fusion for RGB-D Salient Object Detection,” IEEE Access, vol. 7, pp. 55 277–55 284, 2019.

[63] H. Du, Z. Liu, H. Song, L. Mei, and Z. Xu, “Improving rgbd saliency detection using progressive region classification and saliency fusion,” IEEE Access, vol. 4, pp. 8987–8994, 2016.

[64] Y. Niu, Y. Geng, X. Li, and F. Liu, “Leveraging stereopsis for saliency analysis,” in IEEE Conf. Comput. Vis. Pattern Recog., 2012, pp. 454–461.

[65] N. Li, J. Ye, Y. Ji, H. Ling, and J. Yu, “Saliency detection on light field,” in IEEE Conf. Comput. Vis. Pattern Recog., 2014, pp. 2806–2813.

[66] D. Sun, S. Roth, and M. J. Black, “Secrets of optical flow estimation and their principles,” in IEEE Conf. Comput. Vis. Pattern Recog., 2010, pp. 2432–2439.

[67] C. Zhu and G. Li, “A Three-pathway Psychobiological Framework of Salient Object Detection Using Stereoscopic Technology,” in Int. Conf. Comput. Vis. Worksh., 2017, pp. 3008–3014.

[68] R. Cong, J. Lei, H. Fu, W. Lin, Q. Huang, X. Cao, and C. Hou, “An iterative co-saliency framework for rgbd images,” IEEE T. Cybern., vol. 49, no. 1–2, 2017.

[69] R. Cong, J. Lei, H. Fu, Q. Huang, X. Cao, and N. Ling, “Hscs: Hierarchical sparsity based co-saliency detection for rgbd images,” IEEE T. Multimedia, 2018.

[70] 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 T. Image Process., vol. 27, no. 2, pp. 568–579, 2018.

[71] M.-M. Cheng, N. J. Mitrani, X. Huang, P. H. S. Torr, and S.-M. Hu, “Global contrast based salient region detection,” IEEE T. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 569–582, 2015.
