# 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 (D$^3$Net). 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. D$^3$Net 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 D$^3$Net 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 D$^3$Net 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].

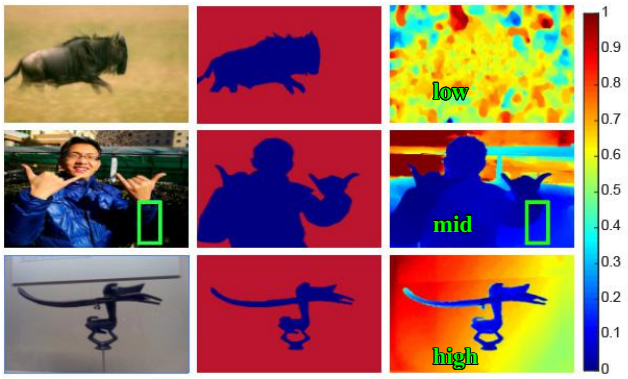



Fig. 1. Left to right: input image, ground truth, and the corresponding depth map. The quality of the depth map from low ($1^{st}$ row), mid ($2^{nd}$ row) to high (last row). As shown in the $2^{nd}$ row, it is difficult to recognize the boundary of the human's arm in the boundary box region. However, it is clearly visible in the depth map. The high-quality depth maps benefit the RGB-D based salient object detection task. These three examples are from the NJU2K [37], our *SIP* and NLPR [39] datasets respectively.

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

Manuscript received July 16, 2019; revised March 9, 2020; accepted May 16, 2020. Date of publication June 3, 2020; date of current version May 3, 2021. This work was supported in part by the Major Project for New Generation of AI under Grant 2018AAA0100400, in part by the NSFC under Grant 61922046, and in part by the Tianjin Natural Science Foundation under Grant 17JCJQJC43700. (*Corresponding author: Ming-Ming Cheng.*)

Deng-Ping Fan was with the College of Computer Science, Nankai University, Tianjin 300350, China. He is now with the Inception Institute of Artificial Intelligence (IIAI), Abu Dhabi, United Arab Emirates.

Zheng Lin, Zhao Zhang, and Ming-Ming Cheng are with the College of Computer Science, Nankai University, Tianjin 300350, China (e-mail: cmm@nankai.edu.cn).

Menglong Zhu is with Google AI, Mountain View, CA 94043 USA.

Color versions of one or more of the figures in this article are available online at https://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2020.2996406

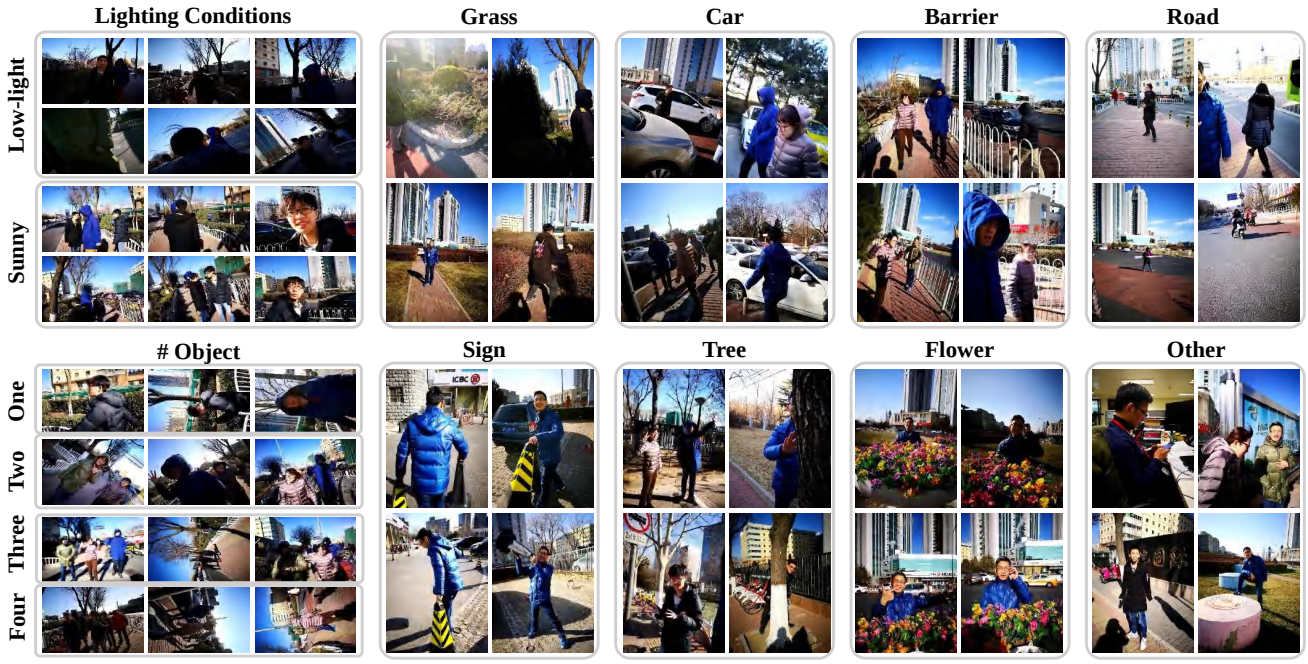



Fig. 2. Representative subsets in our *SIP*. The images in *SIP* are grouped into eight subsets according to background objects (*i.e.*, grass, car, barrier, road, sign, tree, flower, and other), different lighting conditions (*i.e.*, low-light, sunny with clear object boundary) and various number of objects (*i.e.*, 1, 2, $\geq$3).


Fig. 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.

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$\sim$4) have been benchmarked in recent papers [39], [56] (Table II). The 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.

TABLE I
COMPARISON OF CURRENT RGB-D DATASETS IN TERMS OF YEAR (**YEAR**), PUBLICATION (**PUB.**), DATASET SIZE (**DS.**), NUMBER OF OBJECTS IN THE IMAGES (**#OBJ.**), TYPE OF SCENE (**TYPES.**), DEPTH SENSOR (**SENSOR.**), DEPTH QUALITY (**DQ.**, (HIGH-QUALITY DEPTH MAP SUFFERS FROM LESS RANDOM NOISE). SEE LAST ROW IN FIG. 1), ANNOTATION QUALITY (**AQ.**, SEE FIG. 12), WHETHER OR NOT PROVIDE GRAYSCALE IMAGE FROM MONOCULAR CAMERA (**GI.**), CENTER BIAS (**CB.**, SEE FIG. 4 (A)-(B)), AND RESOLUTION (IN PIXEL). H & W DENOTE THE HEIGHT AND WIDTH OF THE IMAGE, RESPECTIVELY.

| No. | Dataset | Year | Pub. | DS. | #Obj. | Types. | Sensor. | DQ. | AQ. | GI. | CB. | Resolution (H×W) |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | *STERE* [63] | 2012 | CVPR | 1K | ∼one | internet | Stereo camera+sift flow [54] | | High | No | High | [251∼1200]×[222∼900] |
| 2 | *GIT* [36] | 2013 | BMVC | 0.08K | multiple | home environment | Microsoft Kinect [52] | | High | No | Low | 640 × 480 |
| 3 | *LFSD* [64] | 2014 | CVPR | 0.1K | one | 60 indoor/40 outdoor | Lytro Illum camera [53] | | High | No | High | 360 × 360 |
| 4 | *DES* [38] | 2014 | ICIMCS | 0.135K | one | 135 indoor | Microsoft Kinect [52] | High | | No | High | 640 × 480 |
| 5 | *NLPR* [39] | 2014 | ECCV | 1K | multiple | indoor/outdoor | Microsoft Kinect [52] | High | | No | High | 640 × 480, 480 × 640 |
| 6 | *NJU2K* [37] | 2014 | ICIP | 1.985K | ∼one | 3D movie/internet/photo | FujiW3 camera+optical flow [65] | | High | No | High | [231∼1213]×[274∼828] |
| 7 | *SSD* [66] | 2017 | ICCVW | 0.08K | multiple | three stereo movies | Sun's optical flow [65] | | | No | Low | 960 × 1080 |
| 8 | ***SIP (Ours)*** | **2020** | TNNLS | **0.929K** | **multiple** | **person in the wild** | **Huawei Mate10** | **High** | **High** | **Yes** | **Low** | **992×744** |

### A. Contribution

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

(1) We have built a new **S**ali**e**nt **P**erson (***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 (**D$^3$Net**), which learns to automatically discard low-quality depth maps using a novel depth depurator unit (DDU). Thanks to the gate connection mechanism, our D$^3$Net can predict salient objects accurately. Extensive experiments demonstrate that our D$^3$Net 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 D$^3$Net 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 D$^3$Net model. Further, in § V-D, we provide a performance comparison between traditional and deep models. We also discuss the experimental results in more depth. 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 D$^3$Net in a background changing application. To better understand the contributions of DDU in the proposed D$^3$Net, in § VII, we present the upper and lower bound of the DDU. All in all, the extensive experimental results clearly demonstrate that our D$^3$Net 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 thus, 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 ($1^{st}$ row in Fig. 1). Thus, we instead design a simple general framework D$^3$Net, 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 objecy 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 cameras 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.

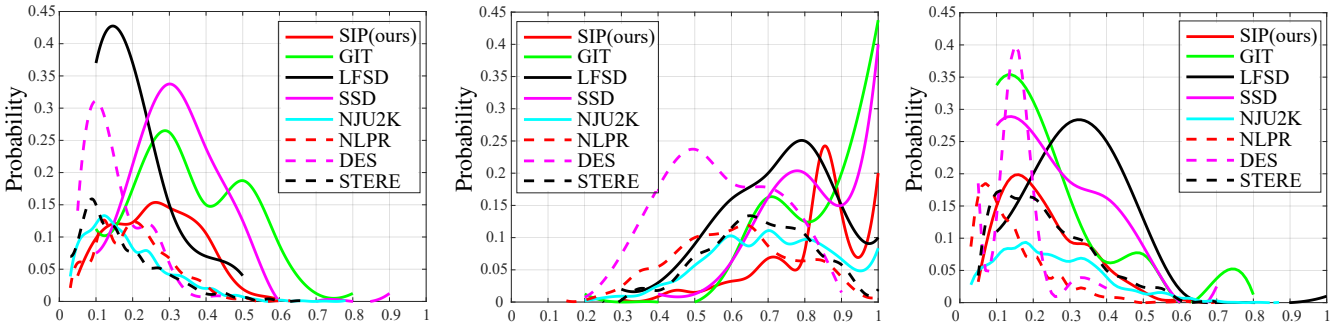
### C. Dataset Statistics

**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

TABLE II

COMPARISON OF 31 CLASSICAL RGB-D BASED SOD ALGORITHMS AND THE PROPOSED BASELINE (D$^3$NET). **TRAIN/VAL SET. (#)** = TRAINING OR VALIDATION SET: NLR = NLPR [39]. NJU = NJU2K [37]. MK = MSRA10K [70]. O = MK + DUTS [71]. **BASIC:** 4PRIORS = 4 PRIORS (REGION, BACKGROUND, DEPTH, AND SURFACE ORIENTATION PRIOR). IPT = INITIALIZATION PARAMETERS TRANSFER. LGBS PRIORS = LOCAL CONTRAST, GLOBAL CONTRAST, 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↑ [59] |
|---|---|---|---|---|---|---|---|---|---|
| 1 | LS [36] | 2013 | BMVC | Without training dataset | One | Markov Random Field | T | ✓ | Not Available |
| 2 | RC [73] | 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 | ICIP | Without training dataset | One | Difference of Gaussian | T | ✓ | 0.780∼0.850 |
| 6 | SRDS [75] | 2014 | DSP | Without training dataset | One | Weighted Color Contrast | T | | Not available |
| 7 | GP [40] | 2015 | CVPRW | Without training dataset | 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 [41] | 2016 | CVPR | Without training dataset | Two | Angular Density Component | T | ✓ | 0.736∼0.890 |
| 10 | DCMC [55] | 2016 | SPL | Without training dataset | Two | Depth Confidence, Compactness, Graph | T | ✓ | 0.743∼0.856 |
| 11 | SE [42] | 2016 | ICME | Without training dataset | Two | Cellular Automata | T | ✓ | 0.771∼0.856 |
| 12 | MCLP [67] | 2017 | Cybernetic | Without training dataset | Two | Addition, Deletion and Iteration Scheme | T | ✓ | Not available |
| 13 | TPF [66] | 2017 | ICCVW | Without training dataset | Four | Cellular Automata, Optical Flow | T | ✓ | Not available |
| 14 | CDCP [46] | 2017 | ICCVW | Without training dataset | Two | Center-dark Channel Prior | T | ✓ | 0.700∼0.820 |
| 15 | DF [44] | 2017 | TIP | NLR (0.75K) + NJU (1.0K) | Three | Laplacian Propagation, LGBS Priors | D | ✓ | 0.759∼0.880 |
| 16 | BED [76] | 2017 | ICCVW | NLR (0.80K) + NJU (1.6K) + MK (9K) | Two | Background Enclosure Distribution | D | ✓ | Not available |
| 17 | MDSF [45] | 2017 | TIP | NLR (0.50K) + NJU (0.5K) | Two | SVM [74], RFR, Ultrametric Contour Map | T | | 0.779∼0.885 |
| 18 | MFF [77] | 2017 | SPL | Without training dataset | One | Minimum Barrier Distance, 3D prior | T | | Not available |
| 19 | Review [56] | 2018 | TCSVT | Without training dataset | Two | Without model introduced | T | | Not available |
| 20 | HSCS [68] | 2018 | TMM | Without training dataset | Two | Hierarchical Sparsity, Energy Function | T | ✓ | Not available |
| 21 | ICS [69] | 2018 | TIP | Without training dataset | One | MCFM, CLP | T | ✓ | Not available |
| 22 | CDB [47] | 2018 | NC | Without training dataset | One | Background Prior | T | ✓ | 0.698∼0.830 |
| 23 | SCDL [78] | 2018 | DSP | NLR (0.75K) + NJU (1.0K) | Two | Silhouette Feature, Spatial Coherence Loss | D | | Not available |
| 24 | PCF [49] | 2018 | CVPR | NLR (0.70K) + NJU (1.5K) | Three | Complementarity-Aware Fusion module [49] | D | | 0.827∼0.925 |
| 25 | CTMF [43] | 2018 | Cybernetic | NLR (0.65K) + NJU (1.4K) | Four | HHA [79], IPT, Hidden Structure Transfer | D | | 0.829∼0.932 |
| 26 | ACCF [80] | 2018 | IROS | NLR (0.65K) + NJU (1.4K) | Three | Attention-Aware | D | | Not available |
| 27 | PDNet [48] | 2019 | ICME | NLR (0.50K) + NJU (1.5K) + O (21K) | Five | Depth-Enhanced Net [48] | D | | Not available |
| 28 | AFNet [61] | 2019 | Access | NLR (0.70K) + NJU (1.5K) | Three | Switch map, Edge-Aware loss | D | | 0.807∼0.887 |
| 29 | MMCI [81] | 2019 | PR | NLR (0.70K) + NJU (1.5K) | Three | HHA [79], Dilated Convolutional | D | | 0.839∼0.928 |
| 30 | TANet [82] | 2019 | TIP | NLR (0.70K) + NJU (1.5K) | Three | Attention-Aware Multi-Modal Fusion | D | | 0.847∼0.941 |
| 31 | CPFP [51] | 2019 | CVPR | NLR (0.70K) + NJU (1.5K) | Five | Contrast Prior, Fluid Pyramid | D | | 0.852∼0.932 |
| 32 | **D$^3$Net (Ours)** | 2020 | | NLR (0.70K) + NJU (1.5K) | Seven | Depth Depurator Unit | D | | 0.862∼0.953 |




(a) Object center to image center distance (b) Margin to image center distance (c) Normalized object size

Fig. 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) | Background Objects | | | | | | | | Object Boundary | | # Object | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | car | flower | grass | road | tree | signs | barrier | other | dark | clear | 1 | 2 | ≥3 |
| #Img | 107 | 9 | 154 | 140 | 97 | 25 | 366 | 32 | 162 | 767 | 591 | 159 | 179 |

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 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 D$^3$Net 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 to achieve robust performance and high generalization ability on various challenging datasets.

### 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 D$^3$Net baseline to efficiently extract the features in a pyramid manner. The entire D$^3$Net 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 224×224 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 × 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 1×1 Conv operation to reduce the channel. For example, let $\mathbf{I_{rgbd}} \in \mathbb{R}^{W \times H \times 4}$ 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 $C_i \times W_i \times H_i$, *i.e.*, $\{64 \times 224 \times 224, 128 \times 112 \times 112, 256 \times 56 \times 56, 512 \times 28 \times 28, 512 \times 14 \times 14, 32 \times 7 \times 7, 32 \times 14 \times 14, 32 \times 28 \times 28, 32 \times 56 \times 56, 32 \times 112 \times 112, 32 \times 224 \times 224\}$ on $\{F_i, \text{i} \in [1,11]\}$, respectively. Note that the $\{F_1, F_2, F_3, F_4, F_5\}$ are corresponding to the five convoluational stages of VGG-16 (*i.e.*, $\{C_1, C_2, C_3, C_4, C_5\}$).

### B. Depth Depurator Unit (DDU)

In the test phase, we further adopt 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 (2$^{nd}$ 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 $\mathbf{S} \in \{S_{rgb}, S_{rgbd}, S_{depth}\}$, then the goal of DDU is to decide which predicted map $\mathbf{P} \in [0,1]^{W \times H}$ is optimal.

$$\mathbf{P} = F_{ddu}(\{S_{rgb}, S_{rgbd}, S_{depth}\}). \tag{1}$$

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

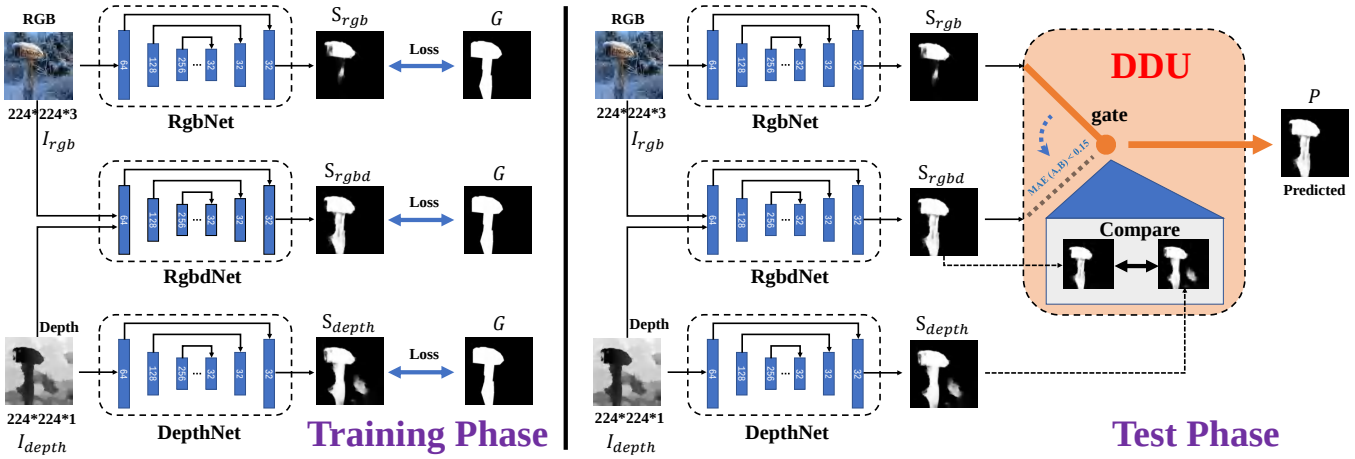



Fig. 5. Illustration of the proposed D$^3$Net. In the training stage (Left), the input RGB and depth images are processed with three parallel sub-networks, *e.g.*, RgbNet, RgbdNet, and DepthNet. The three sub-networks are based on a same modified structure of Feature Pyramid Networks (FPN) (see § IV-A for details). We introduced these sub-networks to obtain three saliency maps (*i.e.*, $S_{rgb}$, $S_{rgbd}$, and $S_{depth}$) which considered both coarse and fine details of the input. In the test phase (Right), a novel depth depurator unit (DDU) (§ IV-B) is utilized for the first time in this work to explicitly discard (*i.e.*, $S_{rgbd}$) or keep (*i.e.*, $S_{rgbd}$) the saliency map introduced by the depth map. In the training/test phase, 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.

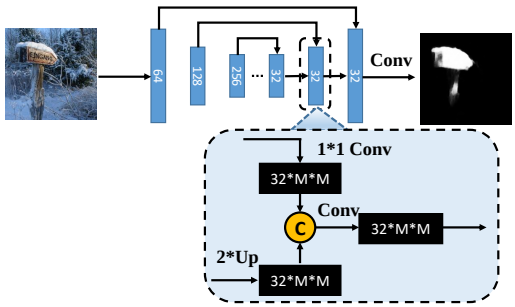
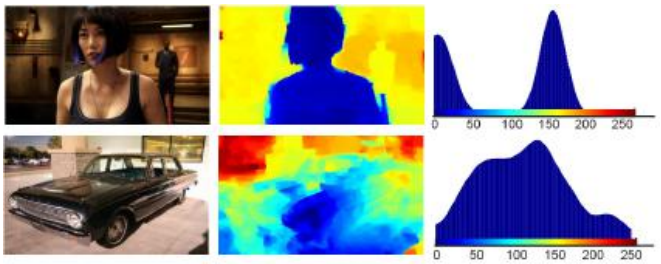



Fig. 6. The Feature Pyramid Network (FPN) is introduced to extract the context-aware information. Different from [97], we further add the sixth layer on the base of VGG-16 and the information merge strategy is concatenation rather than addition. More details can be found in § IV-A.




(a) RGB (b) Depth (c) Histogram

Fig. 7. The smoothed histogram (c) of high-quality ($1^{st}$ row), low-quality ($2^{nd}$ row) depth map, respectively.

is considered in the test phase rather than in the training phase. Specially, a comparison unit $F_{cu}$ is leveraged to assess the similarity between the $S_{depth}$ and $S_{rgbd}$ generated from DepthNet and RgbdNet, respectively.

$$F_{cu} = \begin{cases} 1, & \delta(S_{rgbd}, S_{depth}) \leq t \\ 0, & otherwise, \end{cases} \tag{2}$$

where the $\delta(\cdot)$ represents distance function, and $t$ indicates a fixed threshold. Note that the comparison unit $F_{cu}$ 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 $F_{cu}$ as a gate connection to decide the final/optimal predicted map **P**. Thus, our $F_{ddu}$ module can be formulated as:

$$\mathbf{P} = F_{cu} \cdot S_{rgbd} + \bar{F}_{cu} \cdot S_{rgb}, \tag{3}$$

where $\bar{F}_{cu} = 1 - F_{cu}$. The $F_{cu}$ 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$^3$Net 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 $S_{depth}$ (see first row in Fig. 7). The higher the quality of depth map in $I_{depth}$, the more similarity between the $S_{rgbd}$ and the $S_{depth}$. 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:

$$L(\mathbf{S},\mathbf{G}) = -\frac{1}{N}\sum_{i=1}^{N}\Big(g_i\log(s_i) + (1-g_i)\log(1-s_i)\Big), \tag{4}$$

where $\mathbf{S} \in [0,1]^{224\times224}$ and $\mathbf{G} \in \{0,1\}^{224\times224}$ indicate the estimated saliency map (*i.e.*, $S_{rgb}$, $S_{rgbd}$, or $S_{depth}$) and the GT map, respectively. $g_i \in \mathbf{G}$, $s_i \in \mathbf{S}$, 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 *NLPR* [39] dataset, respectively, as the training data (Please refer to our website for the *Trainlist.txt*). The proposed D$^3$Net 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 $\times$ 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* [64], 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$\times$24 and GTX TITAN X. Since [43], [47], [49], [67]–[69], [80]–[82] have not released their codes, the timings are borrowed from the original papers or provided by the authors. Our D$^3$Net does not apply post-processing (*e.g.*, CRF), thus the computation only takes about 0.015s for a $224 \times 224$ image.

### B. Evaluation Metrics

**MAE $M$.** 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:

$$\text{MAE} = \frac{1}{N}|Sal - G|, \tag{5}$$

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:

$$S_\alpha = \alpha * S_o + (1-\alpha) * S_r, \tag{6}$$

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.

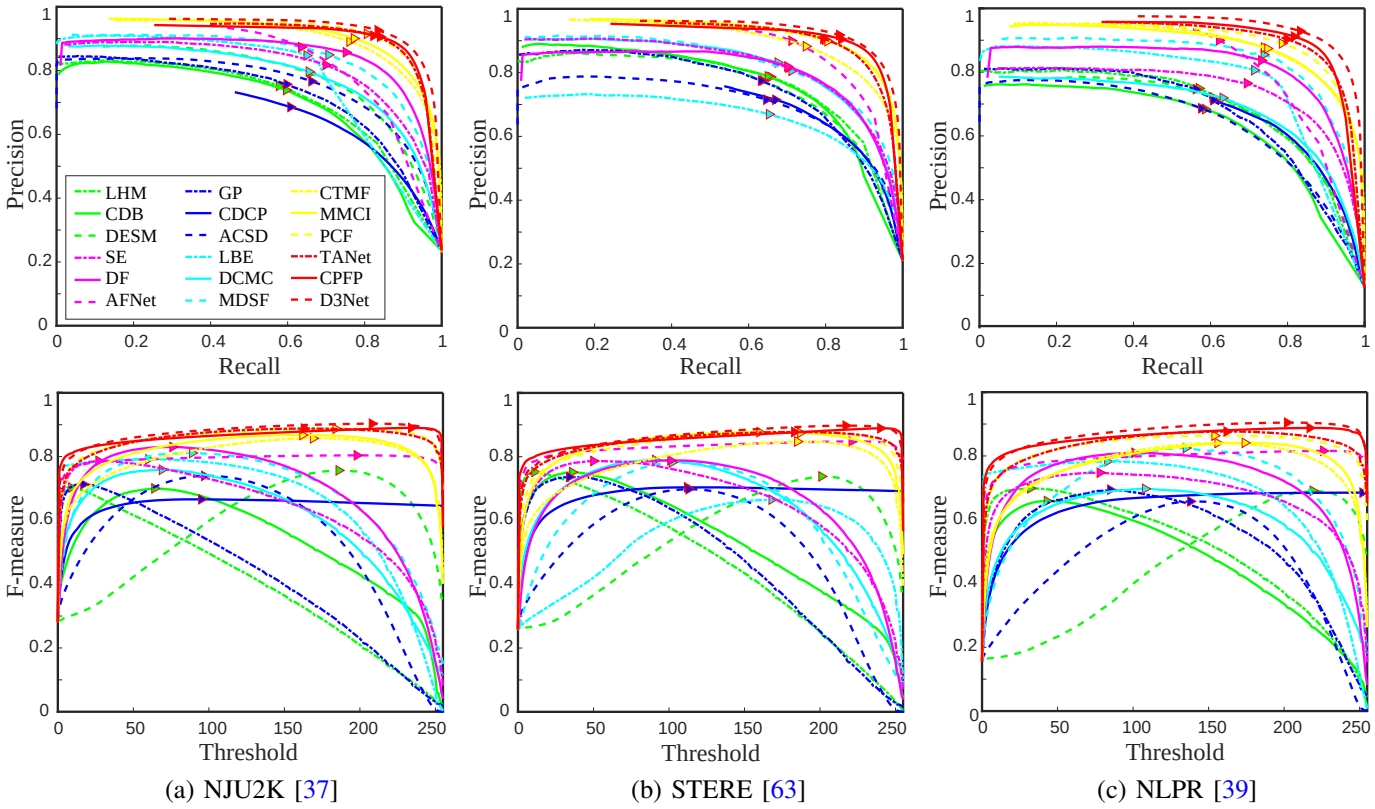
### C. Metric Statistics

For a given metric $\zeta \in \{S_\alpha, F_\beta, E_\xi, M\}$ we consider different statistics. $I_j^i$ denote an image from a specific dataset $D_i$. Thus, $D_i = \{I_1^i, I_2^i, \ldots, I_j^i\}$. Let $\overline{\zeta}(I_j^i)$ be the metric score on image $I_j^i$. The *mean* is the average dataset statistic defined as $M_\zeta(D_i) = \frac{1}{|D_i|}\sum\overline{\zeta}(I_j^i)$, 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 D$^3$Net, 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.




(a) NJU2K [37] (b) STERE [63] (c) NLPR [39]

Fig. 8. PR Curve (top) and F-measures (bottom) for 18 methods on NJU2K, STERE, and NLPR datasets, using various fixed thresholds.

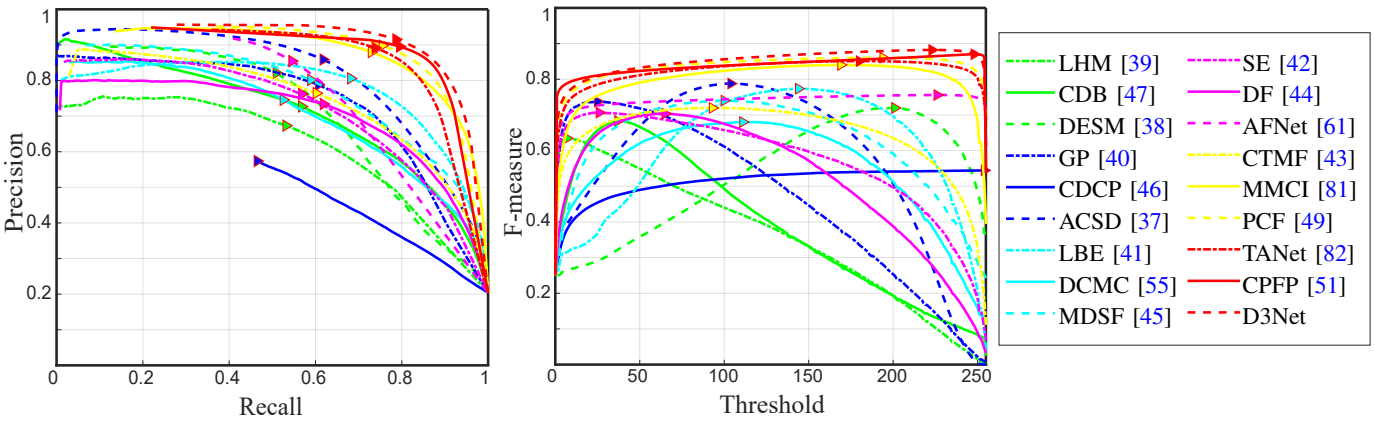



Fig. 9. PR Curve (Left) and F-measures (Right) under different thresholds on the proposed SIP dataset.

**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 D$^3$Net 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% $\sim$ 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 D$^3$Net. 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 $1^{st}$, $4^{th}$, and $5^{th}$ rows) or dramatically (*i.e.*, the $2^{nd}$-$3^{rd}$ rows) damaged. Specifically, in the $3^{rd}$ and $5^{th}$ rows, the depth of the salient object is locally connected with background scenes. Also, the $4^{th}$ 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 ($3^{rd}$-$5^{th}$ rows) or the fine details of the salient object are lost($1^{st}$ row) due to the lack of a cross-modality learning ability. In

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 ON 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 [39] | CDB [47] | DESM [38] | GP [40] | CDCP [46] | ACSD [37] | LBE [41] | DCMC [55] | MDSF [45] | SE [42] | DF [44]† | AFNet [61]† | CTMF [43]† | MMCI [81]† | PCF [49]† | TANet [82]† | CPFP [51]† | **D$^3$Net Ours**† |
| | Time (s) | 2.130 | - | 7.790 | 12.98 | >60.0 | 0.718 | 3.110 | 1.200 | >60.0 | 1.570 | 10.36 | 0.030 | 0.630 | 0.050 | 0.060 | 0.070 | 0.170 | **0.015** |
| | Code | M | - | M | M&C | M&C | C | M&C | M | C | M&C | M&C | Tf | Caffe | Caffe | Caffe | Caffe | Caffe | Pytorch |
| NJU-T [37] | $S_\alpha \uparrow$ | .514 | .624 | .665 | .527 | .669 | .699 | .695 | .686 | .748 | .664 | .763 | .772 | .849 | .858 | .877 | .878 | .879 | **.900** |
| | $F_\beta \uparrow$ | .632 | .648 | .717 | .647 | .621 | .711 | .748 | .715 | .775 | .748 | .804 | .775 | .845 | .852 | .872 | .874 | .877 | **.900** |
| | $E_\xi \uparrow$ | .724 | .742 | .791 | .703 | .741 | .803 | .803 | .799 | .838 | .813 | .864 | .853 | .913 | .915 | .924 | .925 | .926 | **.950** |
| | $M \downarrow$ | .205 | .203 | .283 | .211 | .180 | .202 | .153 | .172 | .157 | .169 | .141 | .100 | .085 | .079 | .059 | .060 | .053 | **.041** |
| | *Rank* | 17 | 16 | 14 | 17 | 15 | 12 | 10 | 13 | 9 | 11 | 7 | 7 | 6 | 5 | 4 | 3 | 2 | 1 |
| STERE [63] | $S_\alpha \uparrow$ | .562 | .615 | .642 | .588 | .713 | .692 | .660 | .731 | .728 | .708 | .757 | .825 | .848 | .873 | .875 | .871 | .879 | **.899** |
| | $F_\beta \uparrow$ | .683 | .717 | .700 | .671 | .664 | .669 | .633 | .740 | .719 | .755 | .757 | .823 | .831 | .863 | .860 | .861 | .874 | **.891** |
| | $E_\xi \uparrow$ | .771 | .823 | .811 | .743 | .786 | .806 | .787 | .819 | .809 | .846 | .847 | .887 | .912 | .927 | .925 | .923 | .925 | **.938** |
| | $M \downarrow$ | .172 | .166 | .295 | .182 | .149 | .200 | .250 | .148 | .176 | .143 | .141 | .075 | .086 | .068 | .064 | .060 | .051 | **.046** |
| | *Rank* | 16 | 12 | 14 | 18 | 13 | 15 | 17 | 10 | 11 | 9 | 8 | 7 | 6 | 3 | 4 | 5 | 2 | 1 |
| DES [38] | $S_\alpha \uparrow$ | .578 | .645 | .622 | .636 | .709 | .728 | .703 | .707 | .741 | .741 | .752 | .770 | .863 | .848 | .842 | .858 | .872 | **.898** |
| | $F_\beta \uparrow$ | .511 | .723 | .765 | .597 | .631 | .756 | .788 | .666 | .746 | .741 | .766 | .728 | .844 | .822 | .804 | .827 | .846 | **.885** |
| | $E_\xi \uparrow$ | .653 | .830 | .868 | .670 | .811 | .850 | .890 | .773 | .851 | .856 | .870 | .881 | .932 | .928 | .893 | .910 | .923 | **.946** |
| | $M \downarrow$ | .114 | .100 | .299 | .168 | .115 | .169 | .208 | .111 | .122 | .090 | .093 | .068 | .055 | .065 | .049 | .046 | .038 | **.031** |
| | *Rank* | 18 | 13 | 14 | 17 | 16 | 12 | 10 | 15 | 11 | 9 | 7 | 8 | 3 | 5 | 6 | 4 | 2 | 1 |
| NLR-T [39] | $S_\alpha \uparrow$ | .630 | .629 | .572 | .654 | .727 | .673 | .762 | .724 | .805 | .756 | .802 | .799 | .860 | .856 | .874 | .886 | .888 | **.912** |
| | $F_\beta \uparrow$ | .622 | .618 | .640 | .611 | .645 | .607 | .745 | .648 | .793 | .713 | .778 | .771 | .825 | .815 | .841 | .863 | .867 | **.897** |
| | $E_\xi \uparrow$ | .766 | .791 | .805 | .723 | .820 | .780 | .855 | .793 | .885 | .847 | .880 | .879 | .929 | .913 | .925 | .941 | .932 | **.953** |
| | $M \downarrow$ | .108 | .114 | .312 | .146 | .112 | .179 | .081 | .117 | .095 | .091 | .085 | .058 | .056 | .059 | .044 | .041 | .036 | **.030** |
| | *Rank* | 14 | 15 | 16 | 18 | 12 | 17 | 10 | 13 | 7 | 11 | 8 | 8 | 5 | 6 | 4 | 3 | 2 | 1 |
| SSD [66] | $S_\alpha \uparrow$ | .566 | .562 | .602 | .615 | .603 | .675 | .621 | .704 | .673 | .675 | .747 | .714 | .776 | .813 | .841 | .839 | .807 | **.857** |
| | $F_\beta \uparrow$ | .568 | .592 | .680 | .740 | .535 | .682 | .619 | .711 | .703 | .710 | .735 | .687 | .729 | .781 | .807 | .810 | .766 | **.834** |
| | $E_\xi \uparrow$ | .717 | .698 | .769 | .782 | .700 | .785 | .736 | .786 | .779 | .800 | .828 | .807 | .865 | .882 | .894 | .897 | .852 | **.910** |
| | $M \downarrow$ | .195 | .196 | .308 | .180 | .214 | .203 | .278 | .169 | .192 | .165 | .142 | .118 | .099 | .082 | .062 | .063 | .082 | **.058** |
| | *Rank* | 16 | 17 | 15 | 11 | 17 | 13 | 14 | 9 | 12 | 9 | 7 | 8 | 6 | 4 | 2 | 2 | 5 | 1 |
| LFSD [64] | $S_\alpha \uparrow$ | .553 | .515 | .716 | .635 | .712 | .727 | .729 | .753 | .694 | .692 | .783 | .738 | .788 | .787 | .786 | .801 | **.828** | .825 |
| | $F_\beta \uparrow$ | .708 | .677 | .762 | .783 | .702 | .763 | .722 | .817 | .779 | .786 | .813 | .744 | .787 | .771 | .775 | .796 | **.826** | .810 |
| | $E_\xi \uparrow$ | .763 | .766 | .811 | .824 | .780 | .829 | .797 | .856 | .819 | .832 | .857 | .815 | .857 | .839 | .827 | .847 | **.872** | .862 |
| | $M \downarrow$ | .218 | .225 | .253 | .190 | .172 | .195 | .214 | .155 | .197 | .174 | .145 | .133 | .127 | .132 | .119 | .111 | **.088** | .095 |
| | *Rank* | 17 | 18 | 16 | 12 | 15 | 11 | 14 | 6 | 13 | 9 | 5 | 10 | 4 | 7 | 8 | 3 | 1 | 2 |
| SIP (Ours) | $S_\alpha \uparrow$ | .511 | .557 | .616 | .588 | .595 | .732 | .727 | .683 | .717 | .628 | .653 | .720 | .716 | .833 | .842 | .835 | .850 | **.860** |
| | $F_\beta \uparrow$ | .574 | .620 | .669 | .687 | .505 | .763 | .751 | .618 | .698 | .661 | .657 | .712 | .694 | .818 | .838 | .830 | .851 | **.861** |
| | $E_\xi \uparrow$ | .716 | .737 | .770 | .768 | .721 | .838 | .853 | .743 | .798 | .771 | .759 | .819 | .829 | .897 | .901 | .895 | .903 | **.909** |
| | $M \downarrow$ | .184 | .192 | .298 | .173 | .224 | .172 | .200 | .186 | .167 | .164 | .185 | .118 | .139 | .086 | .071 | .075 | .064 | **.063** |
| | *Rank* | 17 | 16 | 14 | 12 | 18 | 6 | 9 | 14 | 10 | 11 | 13 | 7 | 8 | 5 | 3 | 4 | 2 | 1 |
| | *All Rank* | 18 | 17 | 15 | 14 | 16 | 13 | 12 | 11 | 10 | 9 | 7 | 8 | 6 | 5 | 4 | 3 | 2 | 1 |

contrast, our D$^3$Net 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

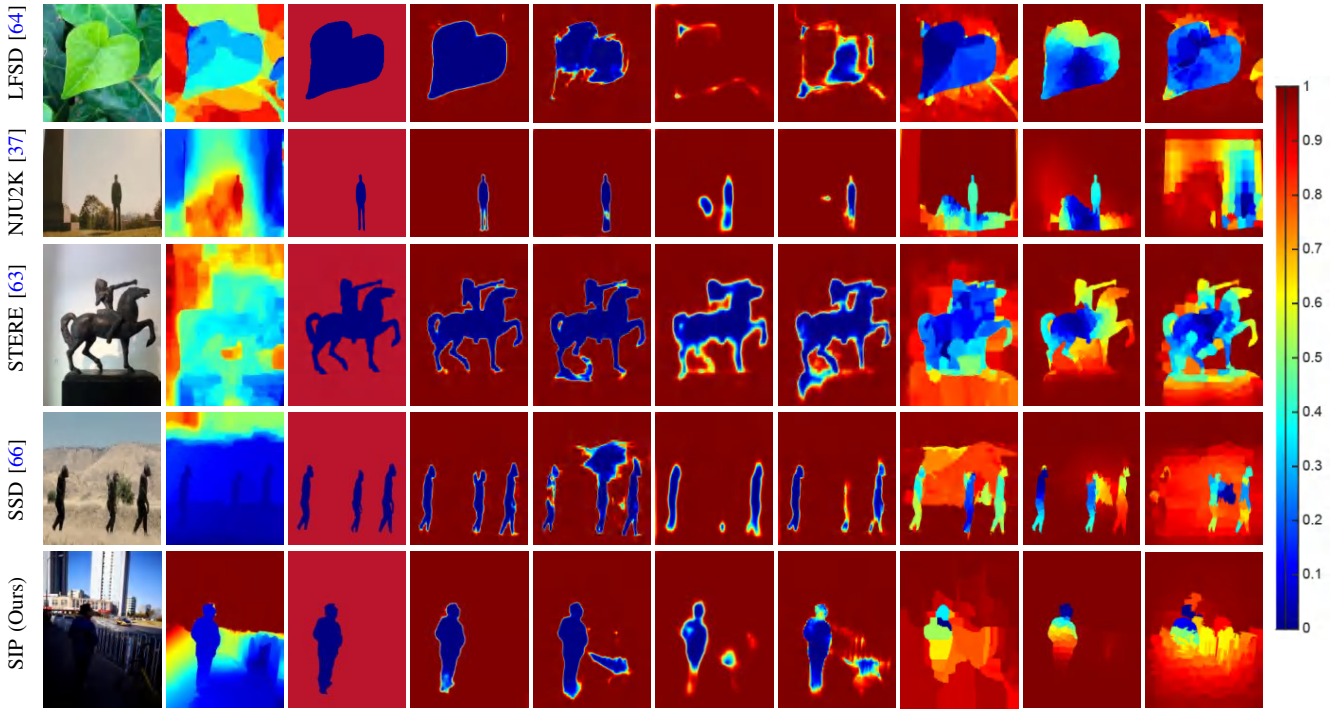



Fig. 10. Visual comparisons with the top-3 CNN-based models (CPFP [51], TANet [82], and PCF [49]) and three classical non-deep methods (MDSF [45], SE [45] and DCMC [55]), on five datasets. Further results can be found in http://dpfan.net/D3NetBenchmark.


(a) Input Image (b) Template (c) Salient object (d) Results

Fig. 11. Examples of book cover maker. See § VI for details.

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 and take certain actions.

### B. Background Changing Application

Background changing techniques have become vital for art designers to leverage the increasing volumes of available image database. Traditional designers utilize photoshop to design their products. This is quite a time-consuming task and requires significant technical knowledge. A large majority of potential users fail to grasp the high-skilled technique in the art design. Thus, an easy-to-use application is needed.

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

TABLE V
S-MEASURE↑ SCORE ON OUR SIP AND THE STERE DATASET. THE SYMBOL ↑ INDICATES THAT THE HIGHER THE SCORE IS, THE BETTER THE MODEL PERFORMS AND VICE VERSA. SEE DETAILS IN § VII.

| Aspects | Model | SIP (Ours) | STERE [63] | DES [38] | LFSD [64] | SSD [66] | NJU2K [37] | NLPR [39] |
|---|---|---|---|---|---|---|---|---|
| w/o DDU | RgbNet | 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 |
| DDU | Lower Bound | 0.822 | 0.881 | 0.870 | 0.788 | 0.817 | 0.875 | 0.897 |
| | **D$^3$Net (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 |

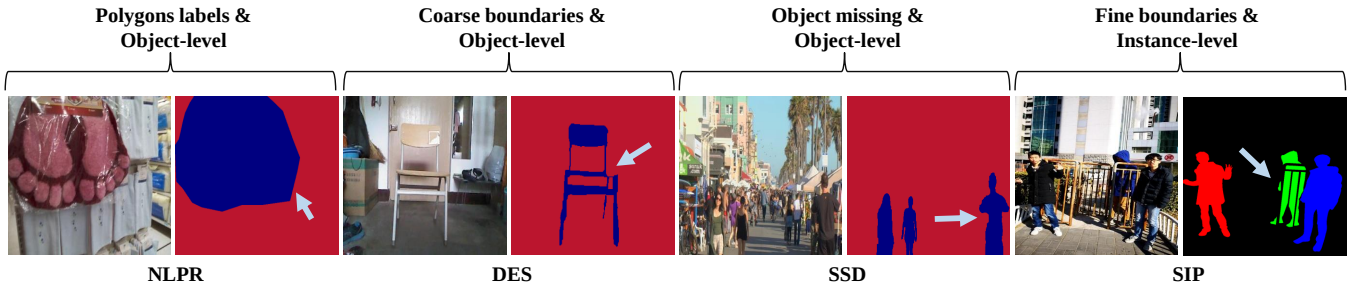



Fig. 12. Comparison with previous object-level datasets, which are labeled with polygons (the foot pad in *NLPR* [39]), coarse boundaries (*i.e.*, the chair in *DES* [38]), and missed object parts (*e.g.*, the person in *SSD* [66]). In contrast, the proposed object-/instance-level SIP dataset is labeled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

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 (RgbNet or RgbdNet). For the SSD dataset, however, the DDU achieves comparable performance to the single stream network (*i.e.*, RgbdNet). It is worth mentioning that D$^3$Net 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 D$^3$Net, we additionally select the optimal path (RgbdNet or RgbNet) of the D$^3$Net. 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 D$^3$Net. From existing results listed in Table V, D$^3$Net still has a $\sim$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 D$^3$Net consists of three sub-networks, which may increase the memory on a light-weight 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 D$^3$Net. (3) Conducting so far the largest-scale ($\sim$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 D$^3$Net 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 D$^3$Net to further boost the performance. After this submission, there are many interesting models, such as UCNet [117], JL-DCF [118], GFNet [119], DMRA [120], ERNet [121], BiANet [122], *etc*, have been released. Please refer to our online leaderboard (http://dpfan.net/d3netbenchmark/) for more details. This website will be updated continually. We foresee this study driving salient object detection towards real-world application scenarios with multiple salient persons and complex interactions through the mobile device (*e.g.*, smartphone or tablets).

**Acknowledgment.** We thank Jia-Xing Zhao, Yun Liu, and Qibin Hou for insightful feedback.

## 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. 3127–3135.

[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. Zhuge, 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. JZhang, 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. Qin, 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. Qin, 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.

[19] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 247–256.

[20] M. A. Islam, M. Kalash, M. Rochan, N. Bruce, and Y. Wang, "Salient object detection using a context-aware refinement network," in *Brit. Mach. Vis. Conf.*, 2017.

[21] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 6609–6617.

[22] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep Image Saliency Computing via Progressive Representation Learning." *IEEE T. Neur. Net. Lear.*, vol. 27, no. 6, pp. 1135–1149, 2016.

[23] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 660–668.

[24] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 1265–1274.

[25] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Eur. Conf. Comput. Vis.* Springer, 2018, pp. 234–250.

[26] Y. Zhuge, Y. Zeng, and H. Lu, "Deep embedding features for salient object detection," in *AAAI Conference on Artificial Intelligence*, 2019.

[27] J. Su, J. Li, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," *arXiv preprint arXiv:1812.10066*, 2018.

[28] P. Jiang, Z. Pan, N. Vasconcelos, B. Cheng, and J. Peng, "Super diffusion for salient object detection," *arXiv preprint arXiv:1811.09038*, 2018.

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

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

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

[32] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, "Contour knowledge transfer for salient object detection," in *Eur. Conf. Comput. Vis.* Springer, 2018, pp. 355–370.

[33] M. Kummerer, T. S. A. Wallis, and M. Bethge, "Saliency benchmarking made easy: Separating models, maps and metrics," in *Eur. Conf. Comput. Vis.* Springer, 2018.

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

[35] M. Amirul Islam, M. Kalash, and N. D. Bruce, "Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 7142–7150.

[36] A. Ciptadi, T. Hermans, and J. M. Rehg, "An in depth view of saliency," in *Brit. Mach. Vis. Conf.*, 2013.

[37] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *IEEE Int. Conf. Image Process.*, 2014, pp. 1115–1119.

[38] Y. Cheng, H. Fu, X. Wei, J. Xiao, and X. Cao, "Depth enhanced saliency detection method," in *Int. Conf. Internet Multi. Comput. Serv.*, 2014, p. 23.

[39] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "Rgbd salient object detection: a benchmark and algorithms," in *Eur. Conf. Comput. Vis.* Springer, 2014, pp. 92–109.

[40] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Ying Yang, "Exploiting Global Priors for RGB-D Saliency Detection," in *IEEE Conf. Comput. Vis. Pattern Recog. Worksh.*, 2015, pp. 25–32.

[41] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for RGB-D salient object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2016, pp. 2343–2350.

[42] J. Guo, T. Ren, and J. Bei, "Salient object detection for rgb-d image via saliency evolution," in *Int. Conf. Multimedia and Expo*, 2016, pp. 1–6.

[43] 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 T. Cybern.*, 2018.

[44] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, and Q. Yang, "RGBD salient object detection via deep fusion," *IEEE T. Image Process.*, vol. 26, no. 5, pp. 2274–2285, 2017.

[45] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, and T. Ren, "Depth-aware salient object detection and segmentation via multiscale discriminative 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 prior," *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. Vis. Comput. 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. Ng, M. Levoy, M. Brédif, G. Duval, M. Horowitz, P. Hanrahan *et al.*, "Light field photography with a hand-held plenoptic camera," *Computer Science Technical Report (CSTR)*, vol. 2, no. 11, pp. 1–11, 2005.

[54] 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.

[55] 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.

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

[57] 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.

[58] 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.

[59] 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.

[60] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 248–255.

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

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.*, no. 99, pp. 1–14, 2017.

[68] 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.

[69] 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.

[70] M.-M. Cheng, N. J. Mitra, 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.

[71] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 136–145.

[72] P. Sauer, T. F. Cootes, and C. J. Taylor, "Accurate regression procedures for active appearance models." in *Brit. Mach. Vis. Conf.*, 2011, pp. 1–11.

[73] K. Desingh, K. M. Krishna, D. Rajan, and C. Jawahar, "Depth really matters: Improving visual salient region detection with depth." in *Brit. Mach. Vis. Conf.*, 2013.

[74] C.-C. Chang and C.-J. Lin, "LIBSVM: a library for support vector machines," *ACM T. Intel. Syst. Tec.*, vol. 2, no. 3, p. 27, 2011.

[75] X. Fan, Z. Liu, and G. Sun, "Salient region detection for stereoscopic images," in *IEEE Conf. Dig. Sig. Process.*, 2014, pp. 454–458.

[76] 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 *Int. Conf. Comput. Vis. Worksh.*, 2017, pp. 2749–2757.

[77] A. Wang and M. Wang, "RGB-D salient object detection via minimum barrier distance transform and saliency fusion," *IEEE Signal Processing Letters*, vol. 24, no. 5, pp. 663–667, 2017.

[78] P. Huang, C.-H. Shen, and H.-F. Hsiao, "Rgbd salient object detection using spatially coherent deep learning framework," in *IEEE Conf. Dig. Sig. Process.*, 2018, pp. 1–5.

[79] S. Gupta, R. Girshick, P. Arbeláez, and J. Malik, "Learning rich features from RGB-D images for object detection and segmentation," in *Eur. Conf. Comput. Vis.* Springer, 2014, pp. 345–360.

[80] H. Chen, Y.-F. Li, and D. Su, "Attention-aware cross-modal cross-level fusion network for RGB-D salient object detection," in *IEEE Int. Conf. Intell. Rob. Syst.*, 2018, pp. 6821–6826.

[81] 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 Recogn.*, vol. 86, pp. 376–385, 2019.

[82] H. Chen and Y. Li, "Three-stream attention-aware network for RGB-D salient object detection," *IEEE T. Image Process.*, 2019.

[83] D.-P. Fan, J.-J. Liu, S.-H. Gao, Q. Hou, A. Borji, and M.-M. Cheng, "Salient objects in clutter: Bringing salient object detection to the foreground," in *Eur. Conf. Comput. Vis.* Springer, 2018, pp. 1597–1604.

[84] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[85] H. Jiang, M.-M. Cheng, S.-J. Li, A. Borji, and J. Wang, "Joint Salient Object Detection and Existence Prediction," *Front. Comput. Sci.*, 2017.

[86] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2015, pp. 5455–5463.

[87] T. Liu, J. Sun, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2007, pp. 1–8.

[88] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Int. Conf. Comput. Vis.*, vol. 2, 2001, pp. 416–423.

[89] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang, "What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017.

[90] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 1155–1162.

[91] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2014, pp. 280–287.

[92] B. W. Tatler, R. J. Baddeley, and I. D. Gilchrist, "Visual correlates of fixation selection: Effects of scale and time," *Vision research*, vol. 45, no. 5, pp. 643–659, 2005.

[93] E. L. Kaufman, M. W. Lord, T. W. Reese, and J. Volkmann, "The discrimination of visual number," *The American Journal of Psychology*, vol. 62, no. 4, pp. 498–525, 1949.

[94] A. Kirillov, R. Girshick, K. He, and P. Dollár, "Panoptic feature pyramid networks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 6399–6408.

[95] X. Chen, R. Girshick, K. He, and P. Dollár, "Tensormask: A foundation for dense object segmentation," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 2061–2069.

[96] Y. Xiong, R. Liao, H. Zhao, R. Hu, M. Bai, E. Yumer, and R. Urtasun, "Upsnet: A unified panoptic segmentation network," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 8818–8826.

[97] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2017, pp. 2117–2125.

[98] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Int. Conf. Learn. Represent.*, 2015.

[99] S.-H. Gao, M.-M. Cheng, K. Zhao, X.-Y. Zhang, M.-H. Yang, and P. Torr, "Res2net: A new multi-scale backbone architecture," *IEEE T. Pattern Anal. Mach. Intell.*, pp. 1–1, 2020.

[100] T. Zhao and X. Wu, "Pyramid feature attention network for saliency detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019, pp. 3085–3094.

[101] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2012, pp. 733–740.

[102] D. Tsai, M. Flagg, and J. Rehg, "Motion coherent tracking with multi-label mrf optimization, algorithms," in *Brit. Mach. Vis. Conf.*, 2010.

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

[104] X. Yang, T. Mei, Y.-Q. Xu, Y. Rui, and S. Li, "Automatic generation of visual-textual presentation layout," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 12, no. 2, p. 33, 2016.

[105] A. Jahanian, J. Liu, Q. Lin, D. Tretter, E. O'Brien-Strain, S. C. Lee, N. Lyons, and J. Allebach, "Recommendation system for automatic design of magazine covers," in *Proceedings of the 2013 international conference on Intelligent user interfaces*. ACM, 2013, pp. 95–106.

[106] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2018, pp. 4510–4520.

[107] J. Zhang, J. Yu, and D. Tao, "Local deep-feature alignment for unsupervised dimension reduction," *IEEE T. Image Process.*, vol. 27, no. 5, pp. 2420–2432, 2018.

[108] S. Mehta, M. Rastegari, L. Shapiro, and H. Hajishirzi, "Espnetv2: A light-weight, power efficient, and general purpose convolutional neural network," *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[109] G.-Y. Nie, M.-M. Cheng, Y. Liu, Z. Liang, D.-P. Fan, Y. Liu, and Y. Wang, "Multi-level context ultra-aggregation for stereo matching," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2019.

[110] J. Yu, B. Zhang, Z. Kuang, D. Lin, and J. Fan, "iprivacy: image privacy protection by identifying sensitive objects via deep multi-task learning," *IEEE T. Info. Foren. Secur.*, vol. 12, no. 5, pp. 1005–1016, 2016.

[111] J. Shen, X. Dong, J. Peng, X. Jin, L. Shao, and F. Porikli, "Submodular function optimization for motion clustering and image segmentation," *IEEE T. Neur. Net. Lear.*, 2019.

[112] D.-P. Fan, G.-P. Ji, G. Sun, M.-M. Cheng, J. Shen, and L. Shao, "Camouflaged object detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[113] 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 scans," *IEEE T. Med. Imag.*, 2020.

[114] J. Yu, C. Zhu, J. Zhang, Q. Huang, and D. Tao, "Spatial pyramid-enhanced netvlad with weighted triplet loss for place recognition," *IEEE T. Neur. Net. Lear.*, 2019.

[115] J. Yu, M. Tan, H. Zhang, D. Tao, and Y. Rui, "Hierarchical deep click feature prediction for fine-grained image recognition," *IEEE T. Pattern Anal. Mach. Intell.*, 2019.

[116] S. He, C. Han, G. Han, and J. Qin, "Exploring duality in visual question-driven top-down saliency," *IEEE T. Neur. Net. Lear.*, 2019.

[117] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F. Sadat Saleh, T. Zhang, and N. Barnes, "UC-Net: Uncertainty Inspired RGB-D Saliency Detection via Conditional Variational Autoencoders," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

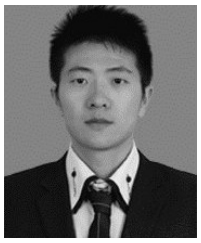
[118] K. F. Fu, D.-P. Fan, G.-P. Ji, and Q. Zhao, "JL-DCF: Joint Learning and Densely-Cooperative Fusion Framework for RGB-D Salient Object Detection," in *IEEE Conf. Comput. Vis. Pattern Recog.*, 2020.

[119] Z. Liu, W. Zhang, and P. Zhao, "A Cross-modal Adaptive Gated Fusion Generative Adversarial Network for RGB-D Salient Object Detection," *Neurocomputing*, 2020.

[120] Y. Piao, W. Ji, J. Li, M. Zhang, and H. Lu, "Depth-induced multi-scale recurrent attention network for saliency detection," in *Int. Conf. Comput. Vis.*, 2019, pp. 7254–7263.

[121] Y. Piao, Z. Rong, M. Zhang, and H. Lu, "Exploit and Replace: An Asymmetrical Two-Stream Architecture for Versatile Light Field Saliency Detection," in *AAAI Conference on Artificial Intelligence*, 2020.

[122] 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.

**Deng-Ping Fan** received his PhD degree from Nankai University of Tianjin in 2019. He joined Inception Institute of Artificial Intelligence (IIAI), UAE in 2019. From 2015 to 2019, he was a Ph.D. candidate in Department of Computer Science, University of Nankai, directed by Prof. Ming-Ming Cheng. He received the Huawei Scholarship in 2017. His current research interests include computer vision, image processing and deep learning.

**Zheng Lin** is currently a Ph.D. candidate with College of Computer Science, Nankai University, under the supervision of Prof. Ming-Ming Cheng. His research interests include deep learning, computer graphics and computer vision.

**Zhao Zhang** received the B.Eng degree from Yangzhou University in 2019. Currently, he is a master student in Nankai University under the supervision of Prof. Ming-Ming Cheng. His research interests includes computer vision and image processing.

**Menglong Zhu** is a Computer Vision Software Engineer at Google. He obtained a Bachelor's degree in Computer Science from Fudan University, in 2010, and a Master's degree in Robotics and a PhD degree in Computer and Information Science from University of Pennsylvania, in 2012 and 2016, respectively. His research interests are on object recognition, 3D object/human pose estimation, human action recognition, visual SLAM and text recognition.

**Ming-Ming Cheng** received his PhD degree from Tsinghua University in 2012. Then he 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 includes computer graphics, computer vision, and image processing. He received research awards including ACM China Rising Star Award, IBM Global SUR Award, CCF-Intel Young Faculty Researcher Program, *etc*.