# Calibrated RGB-D Salient Object Detection

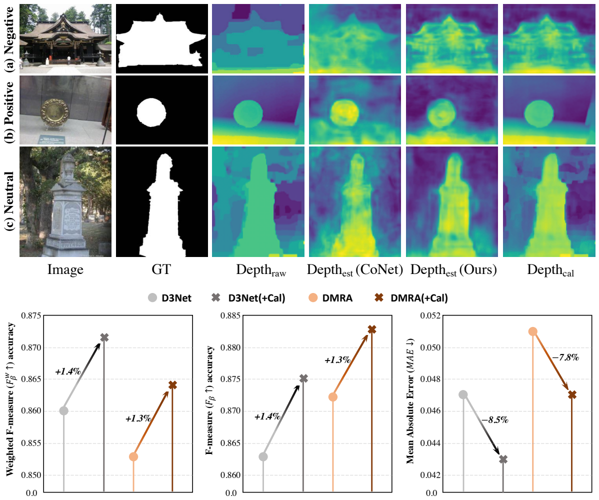
Wei Ji[1,2], Jingjing Li[1], Shuang Yu[2✉], Miao Zhang[3], Yongri Piao[3], Shunyu Yao[3],
Qi Bi[2,4], Kai Ma[2], Yefeng Zheng[2], Huchuan Lu[3], Li Cheng[1✉]
[1]University of Alberta, Canada [2]Tencent Jarvis Lab, Shenzhen, China
[3]Dalian University of Technology, China [4]Wuhan University, Hubei, China
{wji3, jingjin1, lcheng5}@ualberta.ca, {shirlyyu, kylekma, yefengzheng}@tencent.com

## Abstract

*Complex backgrounds and similar appearances between objects and their surroundings are generally recognized as challenging scenarios in Salient Object Detection (SOD). This naturally leads to the incorporation of depth information in addition to the conventional RGB image as input, known as RGB-D SOD or depth-aware SOD. Meanwhile, this emerging line of research has been considerably hindered by the noise and ambiguity that prevail in raw depth images. To address the aforementioned issues, we propose a* ***D****epth* ***C****alibration and* ***F****usion (****DCF****) framework that contains two novel components: 1) a learning strategy to calibrate the latent bias in the original depth maps towards boosting the SOD performance; 2) a simple yet effective cross reference module to fuse features from both RGB and depth modalities. Extensive empirical experiments demonstrate that the proposed approach achieves superior performance against 27 state-of-the-art methods. Moreover, our depth calibration strategy alone can work as a preprocessing step; empirically it results in noticeable improvements when being applied to existing cutting-edge RGB-D SOD models. Source code is available at* *https://github.com/jiwei0921/DCF*.

## 1. Introduction

Salient Object Detection (SOD) is an important computer vision problem that aims to identify and segment the most prominent object in a scene. It has found successful applications in a variety of tasks such as object recognition [59], image retrieval [38, 61], SLAM [37] and video analysis [25, 19, 14]. To tackle the innate challenges in addressing difficult scenes with low texture contrast or in the presence of cluttered backgrounds, depth information has been incorporated as a complementary input source. The growing interests in the development of RGB-D SOD methods [12, 42, 48] are especially boosted by the rapid progress and flourish of varied 3D imaging sensors [29], ranging from the traditional stereo imaging that produces disparity maps, to the more recent structured lighting [76, 30], time-of-flight, light field [63, 71, 72] and LIDAR cameras that directly generate depth images. As showcased by the recent cross-modality fusion schemes [7, 10, 44], adding depth-map on top of RGB image as an extra input leads to superior performance in localizing salient objects on challenging scenes.



(d) Models with Depth$_{raw}$ vs. Depth$_{cal}$ on NJU2K benchmark

Figure 1. **Top**: Examples of different depth qualities; GT denotes the ground-truth saliency map; Depth$_{\mathrm{raw}}$ denotes the original depth map; Depth$_{\mathrm{est}}$ in the $4^{th}$ and $5^{th}$ columns are the estimated depth produced by CoNet [34] and our DCF, respectively; Depth$_{\mathrm{cal}}$ of the last column is generated by our proposed depth calibration strategy. **Bottom**: Accuracy of two representative RGB-D SOD models (D3Net [24] and DMRA [54]) trained with original and calibrated depth ('+Cal'), respectively.

In essence, the actual value of depth in SOD lies in its capability of discerning the object silhouette from background. Nevertheless, practical examination as presented in Fig. 1 implies two main issues that hinder the full exploita-

Wei Ji and Jingjing Li have equal contributions. Wei Ji contributes to this work during internship at Tencent Jarvis Lab.

Shuang Yu and Li Cheng are the corresponding authors.

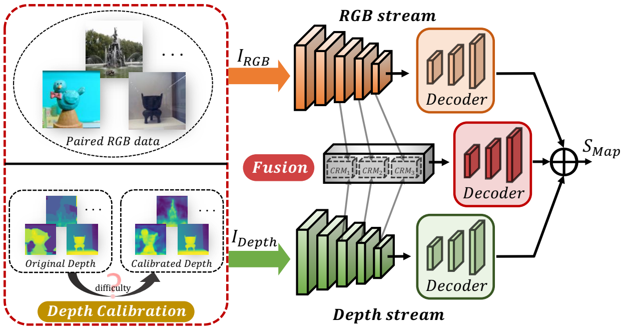


Figure 2. An overview of the proposed **D**epth **C**alibration and **F**usion (DCF) framework.

tion of depth-map: **1)** The depth maps are often exceedingly noisy at the object boundaries, as shown in Fig. 1(a), which may be hampered by the limitation of depth sensors and scene configurations such as occlusion [64], reflection [3, 43] and viewing distance [2]; **2)** Even with correct depth, as exampled by Fig. 1(c), the foreground object often differs only slightly from the surrounding background in the depth maps. This severely limits the potential performance gain of incorporating depth maps compared to using RGB image as the sole input.

To tackle the above two challenges, a **D**epth **C**alibration and **F**usion (DCF) framework is proposed. As illustrated in Fig. 2, our DCF generates an optimal calibration of the depth values that directly promotes salient object detection. Our approach contains the following main contributions:

- A two-step calibration & fusion pipeline is developed: step one involves calibrating the depth image and correcting the latent bias in the original depth maps; step two introduces an effective Cross Reference Module (CRM) to fuse the feature representations from RGB and calibrated depth streams. The performance of the proposed approach is demonstrated with extensive empirical experiments, and compared with 27 state-of-the-art RGB-D SOD methods.
- Our depth calibration module can serve as a preprocessing step that is directly applicable to existing RGB-D SOD methods. By introducing the depth calibration module to the existing RGB-D based SOD methods, the MAE metric of D3Net [24] and DMRA [54] are decreased by 8.5% and 7.8%, respectively, when being evaluated on the widely-used NJU2K benchmark.

## 2. Related Work

Remarkable progress has been evidenced recent years in RGB-image based salient object detection [22, 40, 46, 56, 75, 67, 79]. Meanwhile, the performance still deteriorates severely when objects and their surroundings possess similar appearances, or background scenes are heavily cluttered. As a remedy, the incorporation of depth maps in RGB-D saliency detection has greatly promoted the model performance under those challenging scenarios, benefiting from the embedded rich spatial structure and 3D layout information of the depth maps [5, 15, 17, 18, 21, 41, 62, 77].

The existing RGB-D SOD methods focus more on designing an effective cross-modal fusion strategy to merge the complementary information from RGB and depth channels. Qu et al. [57] attempted to use hand-crafted feature vectors as input to train a CNN-based model and achieved significant improvements over traditional methods [16, 27, 53, 58]. Chen et al. [8] designed a progressive two-stream network, in which the cross-modal residual functions and complementarity-aware supervisions were used to explore cross-model and cross-level complements. Piao et al. [54] proposed a depth-induced multi-scale recurrent attention network, and designed a depth residual block to integrate cross-modal features. Fu et al. [28] jointly learned RGB and depth inputs to mine useful complementary features through a Siamese network. Fan et al. [26] introduced a depth-enhanced module to excavate informative geometric cues from depth features, and also designed a cascaded refinement network to fuse multi-modal and multi-level RGB-D features. To effectively learn the discriminative fused features, Li et al. [44] proposed a cross-modal weighting strategy to encourage comprehensive interactions between RGB and depth information. More detailed descriptions of research in SOD field can be approached in related comprehensive surveys [4, 24, 35, 53, 82].

However, depth maps are occasionally of low quality [52, 76] and thus may contain a lot of noise and misleading information, which results in the performance bottleneck of RGB-D saliency models to certain extent. Recently, there have been several emerging research works shedding light on the influence of unreliable depth and trying to address it. Zhao et al. [78] adopted a contrast prior loss to enhance the color difference between foreground and background of depth data. Similarly, Zhang et al. [69] proposed a semantic guided depth correction subnetwork to produce enhanced depth cues under the assumption that edges of depth map should be aligned with edges of the RGB image. Fan et al. [24] designed a three-stream feature learning network, and performed a depth depurator unit to filter low-quality depth maps during the test phase. Furthermore, Chen et al. [6] leveraged the retrieval of a small set of similar images from external datasets to acquire additional enhanced depth information, and employed a selective fusion way to extract hand-crafted saliency clues from the enhanced depth, original depth and RGB image to predict saliency.

In this paper, we will systematically address the depth-related side effects as discussed previously, and propose a depth calibration and fusion (DCF) framework to tackle this significant challenge. Different from existing methods, our work aims to directly calibrate the raw depth, and the cal-

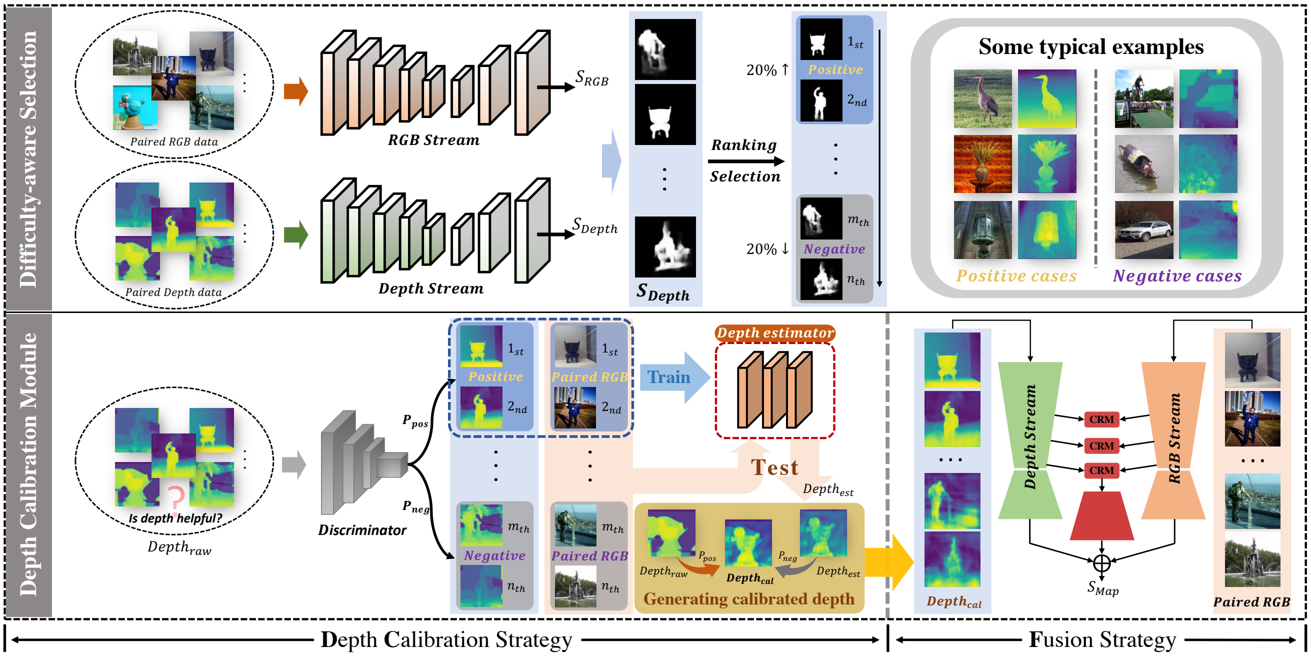


Figure 3. Detailed architecture of our **D**epth **C**alibration and **F**usion (**DCF**) network.

ibrated depth provides more reliable complementary information for saliency models, which significantly boosts the SOD performance. Meanwhile, when directly applying the calibrated depth to existing RGB-D saliency models, noticeable performance gain is also observed.

## 3. Methodology

In this section, we first illustrate the overall architecture of the proposed DCF framework and introduce the key component Depth Calibration (DC) strategy in detail. Additionally, an effective Cross Reference Module (CRM) is proposed to fuse the useful complementary information from both RGB and depth modalities.

### 3.1. Overview

Fig. 2 provides an overview of the proposed DCF framework. Based on a two-stream feature extraction network, it contains two core components: depth calibration and fusion strategies. As presented in Fig. 2 and Fig. 3, a depth calibration (DC) strategy is proposed to correct potential noise caused by unreliable raw depth maps and obtain the calibrated depth $I_{depth}$ (or $Depth_{cal}$). As for the examples shown in Fig. 4, the calibrated depth can manifest the scene layout and identify foreground regions better than the original depth. Now, given the calibrated RGB-D paired data, RGB image $I_{RGB}$ and the calibrated depth $I_{depth}$ are fed into a two-stream feature extraction network to generate hierarchical features. For each stream, an encoder-decoder net [66] is adopted as the backbone. This is followed by a fusion strategy: cross reference modules (CRMs) are designed to integrate the valuable cues from both RGB features and depth features into the cross-modal fused features; this leads to three decoding branches that deal with RGB, depth and fused hierarchical features, respectively. Those features are separately processed and the corresponding outputs are summed up to obtain the final saliency map $S_{Map}$.

### 3.2. Depth Calibration

Effective spatial information from depth map plays an essential role in assisting the localization of salient regions on challenging scenes such as cluttered backgrounds and low-contrast situations. However, unreliable raw depth and potential depth acquisition errors resulted by viewing distance, occlusion or reflection, will impede the model from extracting accurate information from the depth maps.

In order to tackle the performance bottleneck resulted by noisy depth maps, we attempt to calibrate the raw depth to better express the scene layout. There are two key issues that need to be addressed: **1)** How can the model learn to distinguish depth maps with bad quality (negative cases) from the good quality ones (positive cases)? **2)** How to produce the calibrated/corrected depth maps that can both preserve helpful cues from good quality depth maps and correct unreliable information from the bad quality depth maps? Hence, we design the Depth Calibration (DC) strategy, which is the core component of our DCF, as shown in Fig. 3. Two sequential stages are involved to select the representative samples, and generate the calibrated depth maps.

#### 3.2.1 Difficulty-aware Selection Strategy

A difficulty-aware selection strategy is proposed to solve the first key problem. As shown in Fig. 3, it aims to select the most typical negative and positive samples in the training database. These samples are then used to train a discriminator/classifier in predicting the quality of the depth maps, reflecting the reliabilities of depth maps.

We first pre-train two baseline models with the same architecture for RGB data and depth data individually as input under the supervision of saliency ground-truths, denoted as $\psi^{RGB}(\cdot)$, $\psi^{Depth}(\cdot)$, respectively. Then, a selection scheme is designed to measure whether a depth map is able to provide reliable information based on the saliences predicted by the two baseline models. Specifically, according to saliency results generated by the RGB stream and depth stream, we first compute the intersection over union ($IoU$) metric between the predicted saliency and the ground-truth saliency for the two streams, denoted as $IoU_{depth}$ and $IoU_{RGB}$, respectively, for each training sample. Then, the $IoU_{depth}$ scores for all the training samples will be sequentially sorted from large to small. Based on the ranking orders, the samples ranked top 20% of all the training samples will be regarded as typical positive set $\mathcal{P}_{set}$ (i.e., the quality of depth map is acceptable) and the bottom 20% will be regarded as typical negative set $\mathcal{N}_{set}$ (i.e., the quality of depth map is bad and unacceptable). In addition, when $IoU_{depth} > IoU_{RGB}$, these samples will be regarded as positive samples as well, which indicates that raw depth data provides richer global cues to identify foreground regions than RGB input. Some typical examples of both positive cases and negative cases are shown in the upper right corner of Fig. 3.

#### 3.2.2 Depth Calibration Module

Based on the selected representative positive and negative samples, a ResNet-18 [32] based binary discriminator/classifier is trained to evaluate the reliability of the depth map. Here, the selected positive set and negative set are used for the training of the discriminator, $\{Depth_{raw}, 1\} \in \mathcal{P}_{set}$ and $\{Depth_{raw}, 0\} \in \mathcal{N}_{set}$. Our trained discriminator thus is capable of predicting a reliability score $P_{pos}$, indicating the probabilities of the depth map being positive or negative, respectively. The higher $P_{pos}$ is, the better quality the original depth maps have.

In addition, a depth estimator is established, which contains several convolutional blocks using the same architecture as that of [34]. The depth estimator is trained with the RGB image and the good quality depth data pairs from the positive set, i.e., $\{I_{RGB}, Depth_{raw}\} \in \mathcal{P}_{set}$, so as to mitigate the inherent noise resulted by inaccurate raw depth data. In the depth calibration module, instead of directly using the raw depth map which might be unreliable, we replace the original depth map with the weighted summation between the raw depth map and the estimated depth, and the weight is determined by the reliability probability $P_{pos}$ predicted by the discriminator. Thereby, we obtain the calibrated depth map $Depth_{cal}$, as in:

$$Depth_{cal} = Depth_{raw} * P_{pos} + Depth_{est} * (1 - P_{pos}), \quad (1)$$

where $Depth_{est}$ and $Depth_{raw}$ represent the estimated depth from depth estimator and raw depth map, respectively. For better understanding, we visualize the intermediate results of the depth calibration procedure in Fig. 4. For the negative cases with bad quality depth, as seen in the $1^{st}$ and $3^{rd}$ rows in Fig. 4, $Depth_{cal}$ provides more reliable 3D layout information than $Depth_{raw}$. In terms of low-contrast depth data (as seen in the $2^{nd}$ and $5^{th}$ rows), our $Depth_{cal}$ can better manifest the complete scene structure compared to the original depth. As for the original depth map with good quality (as seen in the $4^{th}$ row), the estimated depth $Depth_{est}$ has suboptimal performance compared to $Depth_{raw}$. However, as the the reliability probability $P_{pos}$ predicted by the discriminator will be high for the good quality $Depth_{raw}$, our framework will still be able to obtain a high-quality calibrated depth $Depth_{cal}$.

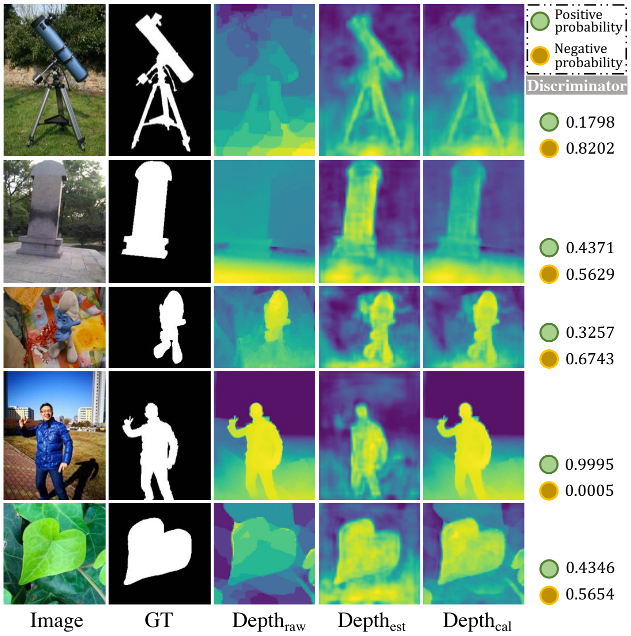


Figure 4. The internal inspections of depth calibration: examples of input depth map Depth$_{raw}$, intermediate estimated depth map Depth$_{est}$ and the calibrated depth map Depth$_{cal}$. The green and yellow circles separately represent positive probability and negative probability produced by the discriminator.

### 3.3. Feature Fusion

After the depth calibration procedure, the calibrated depth map $Depth_{cal}$ together with the RGB image are fed to a two-stream feature extraction network to generate hierarchical features, denoted as $\{F_i^{Depth}\}_{i=3}^{5}$ and $\{F_i^{RGB}\}_{i=3}^{5}$, respectively. Note that we preserve the last three convolution blocks with plentiful semantic features and drop the first two convolutional blocks with high resolution to balance the computational cost. Generally, features extracted from the RGB channel contain rich semantic information and textural information; meanwhile, features from the depth channel contain more discriminative scene layout cues, which are complementary to that of the RGB features. In order to integrate the cross-modality information, our fusing strategy named Cross Reference Module (CRM), is designed and illustrated in Fig. 5.

The proposed CRM aims to mine and combine the most discriminative channels (i.e., feature detectors [65, 68]) among depth and RGB features, and generate more informative features. More specifically, given two input features $F_i^{RGB}$ and $F_i^{Depth}$ produced by the $i^{th}$ convolutional block of the RGB stream and depth stream, respectively, we first employ a global average pooling (GAP) to obtain the global statistics in the RGB and depth views. Then, the two feature vectors are separately fed into a fully connected layer (FC) and a softmax activation function $\delta(\cdot)$ to obtain the channel attention vectors $Att_i^{RGB}$ and $Att_i^{Depth}$, reflecting the importance of the RGB features and depth features, respectively. The attention vectors are then applied on the input feature in a channel-wise multiplication manner. In this way, the CRM will explicitly focus on important features and suppress the unnecessary ones for scene understanding. This procedure can be defined as:

$$Att_i = \delta(\mathcal{W}_i * AvgPooling(F_i) + b_i), \tag{2}$$

where $\mathcal{W}_i$ and $b_i$ represent the parameters of the FC layer for the $i^{th}$ features, and $AvgPooling(\cdot)$ denotes the global average pooling operation. Then, the channel enhancing feature $\dot{F}_i = Att_i \otimes F_i$ is generated, where $\otimes$ denotes the channel-wise multiplication.

In addition, the attention vectors $Att_i^{RGB}$ and $Att_i^{Depth}$ are aggregated by the maximum function to preserve the useful feature channels from both the RGB stream and depth stream, which are then fed to the normalization operation $\mathcal{N}(\cdot)$ to normalize the output to the range from 0 to 1. And thus we obtain the cross-referenced channel attention vector $Att_i^{CR}$. This procedure can be defined as:

$$Att_i^{CR} = \mathcal{N}(Max(Att_i^{RGB}, Att_i^{Depth})). \tag{3}$$

Based on the fusion channel attention vector $Att_i^{CR}$, the enhanced features $\tilde{F}_i^{RGB}$ and $\tilde{F}_i^{Depth}$ can be obtained by summing the $\dot{F_i}^{RGB}$ and $\dot{F_i}^{Depth}$ with the $Att_i^{CR}$ enhanced features. The enhanced features from the RGB branch and depth branch are further concatenated and fed to the $1 \times 1$ convolutional layer to generate the cross-modal fused feature $\mathcal{F}_i$. The procedure can be described as:

$$\tilde{F}_i = \dot{F}_i + Att_i^{CR} \otimes F_i, \tag{4}$$

$$\mathcal{F}_i = Conv_{1\times1}(Concat(\tilde{F_i}^{RGB}, \tilde{F_i}^{Depth})). \tag{5}$$

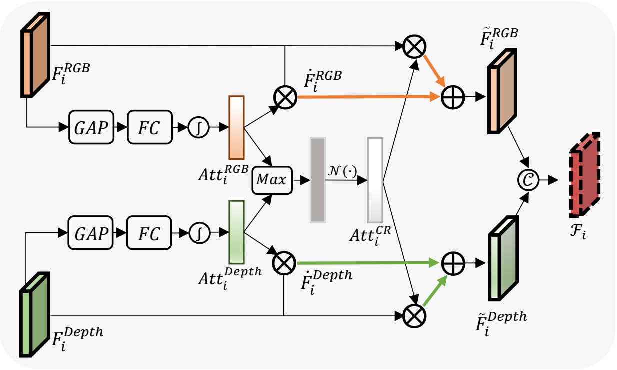


Figure 5. The architecture of the proposed CRM.

Furthermore, a triplet loss is utilized to enhance the obtained cross-modal fused feature $\mathcal{F}_i$, so as to encourage the fused feature to be closer of the foreground, meanwhile enlarging the distance between the foreground feature and the background feature. We use $\mathcal{F}_i$ as the anchor features. Features corresponding to the saliency region are set as the positive, and features of the background region are set as the negative, as in:

$$\mathcal{F}_i^{pos} = \mathcal{F}_i \otimes \mathcal{S}, \tag{6}$$

$$\mathcal{F}_i^{neg} = \mathcal{F}_i \otimes (1 - \mathcal{S}), \tag{7}$$

where $\mathcal{S}$ represents the ground-truth saliency map.

The triplet loss $\mathcal{L}_{triplet}$ then can be calculated as:

$$\mathcal{L}_{triplet} = Max(d(\mathcal{F}_i, \mathcal{F}_i^{pos}) - d(\mathcal{F}_i, \mathcal{F}_i^{neg}) + m, 0), \tag{8}$$

where $d(\cdot)$ indicates the Euclidean distance; $m$ denotes the margin parameter and is set as 1.0 following [60].

After the proposed CRM, we can obtain the cross-modal fused feature $\{\mathcal{F}_i\}_{i=3}^5$, which, together with the original features extracted from the RGB stream $\{F_i^{RGB}\}_{i=3}^5$ and depth stream $\{F_i^{Depth}\}_{i=3}^5$, are further fed to three separate decoders supervised by $\mathcal{S}$, as shown in Fig. 2. Finally, the predictions from three decoders are summed to generate the final saliency map $S_{Map}$.

The optimization objective $\mathcal{L}_{total}$ of the proposed method can be described as:

$$\mathcal{L}_{total} = \mathcal{L}_{RGB} + \mathcal{L}_{Depth} + \mathcal{L}_{fuse} + \frac{\alpha}{N}\sum_{i=3}^{5}\mathcal{L}_{triplet}^{i}, \tag{9}$$

where $\mathcal{L}_{RGB}$, $\mathcal{L}_{Depth}$ and $\mathcal{L}_{fuse}$ denote the binary cross entropy loss between the prediction of each decoder and the ground-truth saliency. $N = 3$ indicates the number of convolutional blocks involved in the triplet loss. In this paper, the hyper-parameter $\alpha$ is set as 0.2 empirically.

## 4. Experiments

### 4.1. Datasets

To evaluate the performance of the proposed DCF framework, we conduct experiments on five representative

Table 1. Quantitative comparison on five representative large-scale benchmark datasets. The best two results are shown in **red** and **blue**, respectively. * means non-deep-learning methods.

| **Pub.** | **Method** | **DUTLF-Depth** [54] | | | | **NJU2K** [36] | | | | **NLPR** [53] | | | | **STERE1000** [50] | | | | **SIP** [24] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $E_\xi$ | $F_\beta^w$ | $F_\beta$ | $MAE$ | $E_\xi$ | $F_\beta^w$ | $F_\beta$ | $MAE$ | $E_\xi$ | $F_\beta^w$ | $F_\beta$ | $MAE$ | $E_\xi$ | $F_\beta^w$ | $F_\beta$ | $MAE$ | $E_\xi$ | $F_\beta^w$ | $F_\beta$ | $MAE$ |
| ICIMCS14 | DES* [13] | .733 | .386 | .668 | .280 | .421 | .241 | .165 | .448 | .735 | .259 | .583 | .301 | .579 | .281 | .594 | .295 | .742 | .352 | .646 | .300 |
| SPL16 | DCMC* [20] | .712 | .290 | .406 | .243 | .796 | .506 | .715 | .167 | .684 | .265 | .328 | .196 | .655 | .551 | .742 | .148 | .787 | .426 | .646 | .186 |
| ECCV14 | LHM* [53] | .767 | .350 | .659 | .174 | .722 | .311 | .625 | .201 | .772 | .320 | .520 | .119 | .484 | .379 | .703 | .172 | .722 | .286 | .593 | .182 |
| CAIP17 | MB* [84] | .691 | .464 | .577 | .156 | .643 | .369 | .492 | .202 | .814 | .574 | .637 | .089 | .693 | .455 | .572 | .178 | .715 | .474 | .573 | .163 |
| TCyb17 | CTMF [31] | **.884** | .690 | **.792** | .097 | .864 | .732 | .788 | .085 | .869 | .691 | .723 | .056 | .841 | .747 | .771 | .086 | .824 | .551 | .684 | .139 |
| TIP17 | DF [57] | .842 | .542 | .748 | .145 | .818 | .552 | .744 | .151 | .838 | .524 | .682 | .099 | .691 | .596 | .742 | .141 | .794 | .411 | .672 | .186 |
| ICCVW17 | CDCP* [85] | .794 | .530 | .633 | .159 | .751 | .522 | .618 | .181 | .785 | .512 | .591 | .114 | .751 | .596 | .666 | .149 | .721 | .411 | .494 | .224 |
| CVPR18 | PCA [8] | .858 | .696 | .760 | .100 | .896 | .811 | .844 | .059 | .916 | .772 | .794 | .044 | .887 | .801 | .826 | .064 | .898 | .777 | .824 | .071 |
| TIP19 | TANet [9] | .866 | **.712** | .779 | **.093** | .893 | .812 | .844 | .061 | .916 | .789 | .795 | .041 | .893 | .804 | .835 | .060 | .893 | .762 | .809 | .075 |
| ICME19 | PDNet [83] | .861 | .650 | .757 | .112 | .890 | .798 | .832 | .062 | .876 | .659 | .740 | .064 | .880 | .799 | .813 | .071 | .802 | .503 | .620 | .166 |
| PR19 | MPCI [10] | .855 | .636 | .753 | .113 | .878 | .749 | .813 | .079 | .871 | .688 | .729 | .059 | .873 | .757 | .829 | .068 | .886 | .726 | .795 | .086 |
| CVPR19 | CPFP [78] | .814 | .644 | .736 | .099 | .895 | .837 | .850 | .053 | .924 | .820 | .822 | .036 | .912 | .808 | .830 | .051 | .899 | .798 | .818 | .064 |
| CVPR20 | JL-DCF [28] | - | - | - | - | - | - | - | - | **.954** | **.882** | .878 | **.022** | .919 | .857 | **.869** | **.040** | **.919** | **.844** | **.873** | **.051** |
| CVPR20 | S2MA [47] | - | - | - | - | - | - | - | - | .938 | .852 | .853 | .030 | .907 | .825 | .855 | .051 | .911 | .825 | .849 | .058 |
| CVPR20 | UCNet [69] | - | - | - | - | - | - | - | - | .953 | .878 | **.890** | .025 | .922 | **.867** | **.885** | **.039** | .913 | .836 | .868 | **.051** |
| TNNLS20 | D3Net [24] | .847 | .668 | .756 | .097 | **.913** | .860 | .863 | **.047** | .943 | .854 | .857 | .030 | .920 | .845 | .855 | .046 | .902 | .808 | .835 | .063 |
| ECCV20 | CMWN [44] | - | - | - | - | .910 | .855 | **.878** | **.047** | .940 | .856 | .859 | .029 | .917 | .847 | **.869** | .043 | .906 | .811 | .851 | .062 |
| ECCV20 | BBSNet [26] | .833 | .663 | .774 | .120 | **.924** | **.884** | **.902** | **.035** | .952 | .879 | .882 | .023 | **.925** | .858 | **.885** | .041 | .916 | .830 | .872 | **.055** |
| **Ours**$_{NJU+NLPR}$ | | **.890** | **.766** | **.804** | **.071** | **.924** | **.893** | **.902** | **.035** | **.957** | **.892** | **.891** | **.021** | **.927** | **.873** | **.885** | **.039** | **.920** | **.848** | **.875** | **.051** |
| ICCV19 | DMRA [54] | .927 | .858 | .883 | .048 | .908 | .853 | .872 | .051 | .942 | .845 | .855 | .031 | .923 | .841 | .876 | .049 | .863 | .750 | .819 | .085 |
| CVPR20 | SSF [73] | .946 | .894 | .914 | .034 | .913 | .871 | .886 | .043 | .949 | .874 | .875 | .026 | .921 | .850 | .867 | .046 | .911 | .829 | .851 | .056 |
| CVPR20 | A2dele [55] | .924 | .864 | .890 | .043 | .897 | .851 | .874 | .051 | .945 | .867 | .878 | .028 | .915 | .855 | .874 | .044 | .892 | .793 | .825 | .070 |
| ACMM20 | FRDT [74] | .941 | .878 | .902 | .039 | .917 | .862 | .879 | .048 | .946 | .863 | .868 | .029 | .925 | .858 | .872 | .042 | .905 | .817 | .854 | .063 |
| ECCV20 | DANet [81] | .925 | .847 | .884 | .047 | - | - | - | - | .949 | .858 | .871 | .028 | .914 | .830 | .858 | .047 | .916 | .829 | .864 | .054 |
| ECCV20 | HDFNet [51] | .934 | .865 | .892 | .040 | .915 | .879 | **.893** | **.038** | .948 | .869 | .878 | .027 | .925 | .863 | .879 | **.040** | **.918** | .835 | .863 | **.051** |
| ECCV20 | CoNet [34] | **.947** | .896 | .908 | .034 | .911 | .856 | .872 | .047 | .934 | .850 | .848 | .031 | **.928** | **.874** | **.885** | **.037** | .909 | .814 | .842 | .063 |
| ECCV20 | PGAR [11] | .944 | .889 | .914 | .035 | .915 | .871 | **.893** | .042 | **.955** | **.881** | **.885** | **.024** | .919 | .856 | .880 | .041 | .908 | .822 | .854 | .055 |
| ECCV20 | ATSA [70] | **.947** | **.901** | **.918** | **.032** | **.921** | **.883** | **.893** | **.040** | .945 | .867 | .876 | .028 | .919 | .866 | .874 | **.040** | .912 | **.848** | **.871** | **.053** |
| **Ours**$_{DUT+NJU+NLPR}$ | | **.952** | **.909** | **.926** | **.030** | **.922** | **.884** | **.897** | **.038** | **.956** | **.892** | **.893** | **.023** | **.931** | **.880** | **.890** | **.037** | **.920** | **.850** | **.877** | **.051** |

large-scale RGB-D SOD datasets, including **DUT-D** [54], **NJU2K** [36], **NLPR** [53], **STERE1000** [50] and **SIP** [24]. **DUT-D** [54] (i.e., DUTLF-Depth) consists of 800 indoor and 400 outdoor scenes images paired with corresponding depth maps. **NJU2K** [36] and **NLPR** [53] contain 1985 and 1000 paired stereo images, respectively. **STERE1000** [50] contains 1000 stereoscopic images downloaded from the Internet. **SIP** [24] is a high-quality RGB-D dataset with 929 samples. Due to the limited space, the results on two more datasets (LFSD [45] and DES [13]) can be accessed in the released github webpage.

To make a fair comparison, we conduct experiments with two different training setups, based on the two current mainstream training settings. For the first training setting, we use 1485 samples from the NJU2K and 700 samples from NLPR as the training set following the same setup as [24, 28, 69]. For the second one, we follow the same training settings as existing works [11, 54, 73], where 800 samples from DUT-D, 1485 samples from NJU2K and 700 samples from NLPR are used the training set. The remaining images and other public datasets are used for testing. To alleviate potential overfitting, images in the training set are augmented with randomly rotating, cropping and flipping.

## 4.2. Evaluation Metrics

Four widely-used metrics are adopted to evaluate the model performance, including E-measure ($E_\xi$) [23], weighed F-measure ($F_\beta^w$) [49], F-measure ($F_\beta$) [1] and Mean Absolute Error ($MAE$) [4]. In addition, we also evaluate the performance of the depth estimator on two RGB-D datasets with high-quality depth maps. The performance of the estimated depth is evaluated with Root Mean Square Error ($RMSE$), absolute relative error ($AbsRel$), squared relative error ($SqRel$) and depth accuracy at various thresholds $1.25$, $1.25^2$ and $1.25^3$, as suggested by [80].

## 4.3. Implementation Details

The framework is implemented with PyTorch and trained using a Tesla P40 GPU with 24GB memory. The backbone network [66] is equipped with ResNet-50 encoder part [32], with initial parameters pre-trained in ImageNet [39]. All images are uniformly resized to the dimension of $352 \times 352$ pixels for training and inferring. The proposed network is trained in a multi-stage manner and it converges after a total of 250 epochs, including 120 epochs for the difficulty-aware selection module in the first stage, 60 epochs for the depth calibration module in the second stage, and 70 epochs for feature fusion in the last stage. During the whole training procedure, the learning rate is set to $1 \times 10^{-4}$, and Adam optimizer is adopted with mini-batch of 16. During the model inference phase, the proposed framework predicts saliency maps in an end-to-end manner and no post-processing procedure (e.g., CRF [33]) is applied in this work.

## 4.4. Comparison with State-of-the-arts

The proposed method is evaluated and compared with 27 RGB-D SOD methods, including 22 deep-learning-based methods and 5 non-deep-learning ones (marked with * in Table 1). For a fair comparison, the results of the compar-

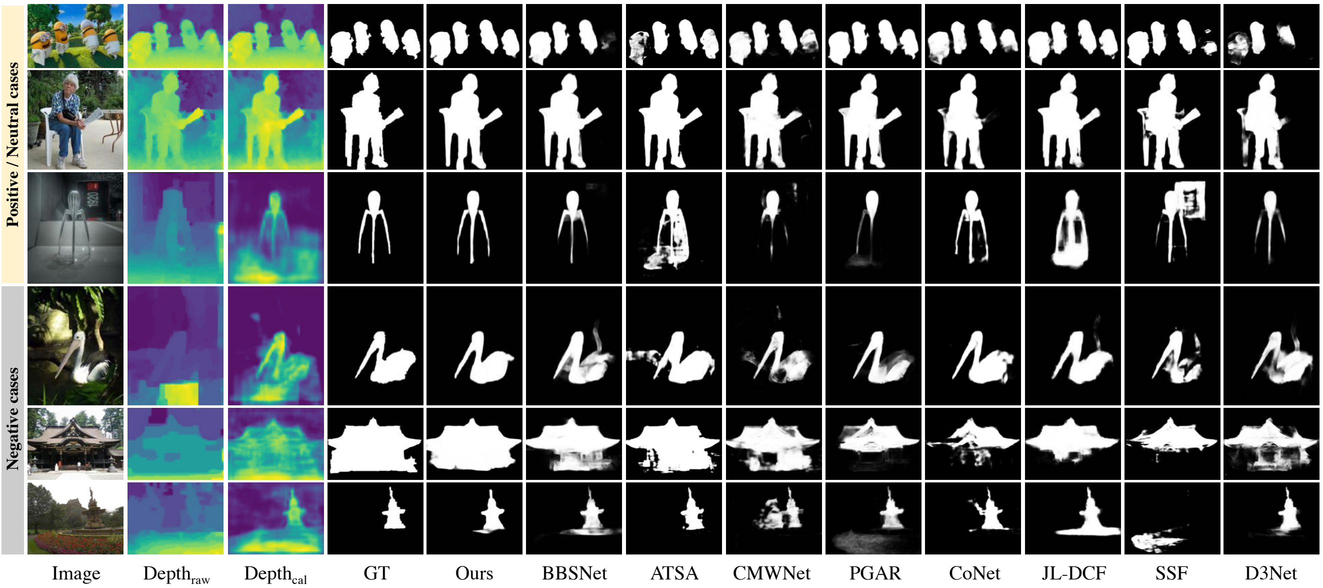


Figure 6. Visual comparisons of the proposed model and existing state-of-the-art algorithms.

ison methods are generated by authorized codes or directly provided by authors.

**Quantitative Evaluation.** Table 1 lists the quantitative comparison results. Following the main-stream training setups as that of [24] and [54], two different training settings are adopted, the results of which are independently listed in the first and second block of Table 1. Overall, our proposed approach achieves superior performance compared to the state-of-the-art methods with both training setups on the five commonly used SOD datasets.

**Qualitative Evaluation.** Fig. 6 shows some representative samples generated by the proposed methods and several top-ranking RGB-D approaches on several challenging cases, including the long distance, cluttered background, sharp boundary and multiple objects. As shown in the third column of Fig. 6, the calibrated depth ($Depth_{cal}$) can provide richer 3D layout cues than the raw depth ($Depth_{raw}$). For the challenging scenes with low-quality depth map resulted by reflection (e.g., the $4^{th}$ row) and viewing distance (e.g., the $5^{th}$ and $6^{th}$ rows), the proposed method can better identify the salient objects by taking advantage of the reliable spatial cues from the calibrated depth map $Depth_{cal}$. Therefore, both quantitative and qualitative evaluations demonstrate the effectiveness of the proposed depth calibration and fusion framework.

## 4.5. Ablation Study

To verify the effectiveness of the proposed modules, ablation studies are performed over each component of the DCF framework to investigate their performance gains.

**RGB Stream vs. Depth Stream.** Table 2 (a) and (b) compare the saliency prediction performance of the baseline models using RGB data as input (RGB stream) and the original depth data as input (depth stream), respectively. The RGB stream achieves better performance than that of the depth stream using original depth maps, indicating that the RGB input contains more semantic and texture information than that of the depth input. In addition, for the SIP dataset with high-quality depth maps, the performance of the depth stream is closer to that of the RGB stream, compared to other datasets with lower-quality depth maps. This again verifies the assumption that reliable depth cues can help the model to identify the salient regions better.

**Effect of depth calibration strategy.** To evaluate the effectiveness of the depth calibration strategy, we first compare the baseline model performance with the original depth as input (depth stream) versus that of using the calibrated depth (calibrated depth stream). As listed in Table 2 (b) and (c), the calibrated depth reduces the $MAE$ metric by averagely 14.51% on four datasets. A relatively smaller performance gain is achieved on the SIP dataset compared with the rest datasets, which is reasonable since high-quality SIP has already provided reliable depth cues in the original depth maps. For better understanding, Fig. 4 visualizes several representative examples of the original depth map, the estimated depth as well as the final calibrated depth map, and as shown in Fig. 7, features map $F^{Depth}_{cal}$ extracted from calibrated depth can capture scene layout information better than $F^{Depth}_{raw}$ from raw depth (see $1^{st}$ vs. $2^{nd}$ rows).

We have also evaluated quality of the estimated depth generated by the depth estimator on two datasets with high-quality depth maps, including the SIP dataset and the DES dataset. As listed in Table 3, our depth estimator achieves more accurate depth estimation, compared with CoNet [34]. Also note that our depth estimator is trained by only 20% of the training set, meanwhile CoNet was trained by 100% of

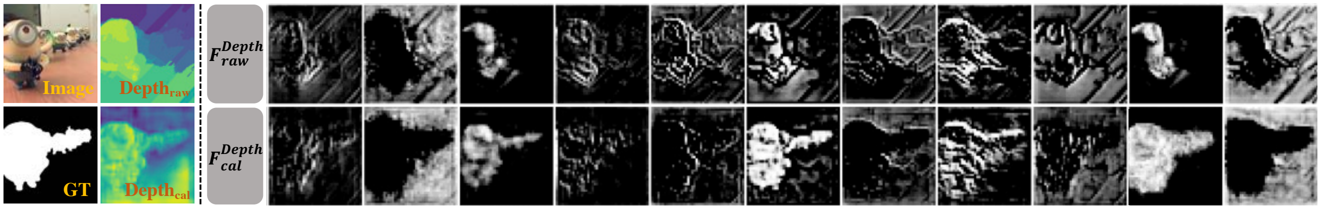


Figure 7. Visualization of feature representation maps in the proposed cross reference module (CRM), where $F_{raw}^{Depth}$ and $F_{cal}^{Depth}$ denote extracted features from backbone with raw depth and calibrated depth as input, respectively. It is observed that the calibrated depth feature maps capture richer structural information than feature maps from raw depth.

Table 2. Quantitative comparison with different ablation settings.

| | | **NJU2K** [36] | | | | **NLPR** [53] | | | | **STERE1000** [50] | | | | **SIP** [24] | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Index.** | **Model.** | $E_\xi$ | $F_\beta^w$ | $F_\beta$ | $MAE$ | $E_\xi$ | $F_\beta^w$ | $F_\beta$ | $MAE$ | $E_\xi$ | $F_\beta^w$ | $F_\beta$ | $MAE$ | $E_\xi$ | $F_\beta^w$ | $F_\beta$ | $MAE$ |
| (a) | RGB Stream | .905 | .866 | .869 | .046 | .942 | .860 | .855 | .028 | .916 | .856 | .863 | .047 | .908 | .813 | .839 | .063 |
| (b) | Depth Stream | .885 | .800 | .831 | .068 | .915 | .794 | .800 | .044 | .823 | .609 | .695 | .122 | .903 | .802 | .845 | .068 |
| (c) | Calibrated Depth Stream | .896 | .824 | .840 | .059 | .925 | .819 | .821 | .039 | .873 | .742 | .778 | .083 | .906 | .804 | .852 | .067 |
| (d) | (a)+(c)+Direct fusion | .910 | .867 | .878 | .043 | .945 | .862 | .859 | .026 | .919 | .863 | .867 | .044 | .913 | .822 | .859 | .060 |
| (e) | (a)+(c)+CRM (w/o $\mathcal{L}_{triplet}$) | .919 | .882 | .890 | .038 | .954 | .887 | .885 | .023 | .921 | .866 | .877 | .042 | .919 | .845 | .869 | .052 |
| (f) | (a)+(c)+CRM (**Ours**) | **.924** | **.893** | **.902** | **.035** | **.957** | **.892** | **.891** | **.021** | **.927** | **.873** | **.885** | **.039** | **.920** | **.848** | **.875** | **.051** |

Table 3. Quantitative comparison with state-of-the-art method CoNet [34] on the accuracy of the estimated depth, evaluating on two high-quality RGB-D datasets SIP [24] and DES [13]. ↑ and ↓ represent high and low scores are better, respectively.

| * | | $RMSE\downarrow$ | $AbsRel\downarrow$ | $SqRel\downarrow$ | $\delta < 1.25\uparrow$ | $\delta < 1.25^2\uparrow$ | $\delta < 1.25^3\uparrow$ |
|---|---|---|---|---|---|---|---|
| SIP [24] | CoNet | 0.4350 | 0.1507 | 0.0947 | 0.6713 | 0.9060 | 0.9846 |
| | Ours | 0.4289 | 0.1482 | 0.0907 | 0.6866 | 0.9168 | 0.9867 |
| DES [13] | CoNet | 0.6426 | 0.2586 | 0.2023 | 0.4364 | 0.7446 | 0.9317 |
| | Ours | 0.4794 | 0.1978 | 0.1192 | 0.5569 | 0.8764 | 0.9851 |

Table 4. Accuracy of the state-of-the-art RGB-D saliency models (i.e., D3Net [24] and DMRA [54]) trained with our calibrated depth vs. the raw depth. '+Cal' represents the models trained on the calibrated depth.

| * | DUT-D [54] | | | NJU2K [36] | | |
|---|---|---|---|---|---|---|
| | $F_\beta^w$ | $F_\beta$ | $MAE$ | $F_\beta^w$ | $F_\beta$ | $MAE$ |
| D3Net [24] | 0.668 | 0.756 | 0.097 | 0.860 | 0.863 | 0.047 |
| **D3Net(+Cal)** | 0.747 | 0.788 | 0.081 | 0.872 | 0.875 | 0.043 |
| DMRA [54] | 0.858 | 0.883 | 0.048 | 0.853 | 0.872 | 0.051 |
| **DMRA(+Cal)** | 0.875 | 0.899 | 0.043 | 0.864 | 0.883 | 0.047 |

the same training set, which also demonstrates the effectiveness of our difficulty-aware selection strategy.

Furthermore, to verify the generalization capability of the proposed depth calibration module, we have also applied the calibrated depth on two state-of-the-art SOD models, including D3Net [24] and DMRA [54]. As listed in Table 4, by replacing the original depth map with the calibrated depth to train D3Net and DMRA, noticeable performance gains have been achieved for the DUT-D dataset and NJU2K dataset. The $MAE$ metric has been decreased by 12.5% and 9.1% for D3Net and DMRA, respectively. Therefore, extensive experiments have demonstrated the advantages of the proposed depth calibration strategy.

**Effect of fusion strategy.** For the cross-modality fusion module to integrate the RGB and depth features, a straightforward solution is to use concatenation followed by convolution operations to fuse the complementary features from RGB and depth (direct fusion). In Table 2, by comparing (d) and (f), we can see that the proposed CRM can better fuse the complementary information from RGB and depth features, compared with direct feature fusion. Meanwhile, compared to (f) in Table 2, i.e., the final framework, when excluding the triplet loss from the framework, performance drop is observed on all the experimental datasets, indicating the effectiveness of the triplet loss in enhancing feature representations. In summary, quantitative and qualitative analysis showed that our DCF framework can effectively capture reliable depth information and integrate complementary cross-modal features.

## 5. Conclusion

In this work, a **D**epth **C**alibration and **F**usion (**DCF**) framework is proposed for accurate RGB-D SOD. Firstly, a depth calibration strategy is designed to correct the potential noise from unreliable raw depth. The calibrated depth has been proved to effectively improve the model performance, for both the proposed framework and state-of-the-art RGB-D saliency models. Additionally, a cross reference module is proposed to effectively integrate the complementary cues from RGB and depth features. Extensive experiments demonstrated the superior performance of our approach over 27 state-of-the-art methods.

**Acknowledgement.** This work was funded by the Key-Area Research and Development Program of Guangdong Province, China (No. 2018B010111001), National Key R&D Program of China (2018YFC2000702) and the Scientific and Technical Innovation 2030-'New Generation Artificial Intelligence' Project (No. 2020AAA0104100), University of Alberta Start-up Grant, UAHJIC grants, and NSERC Discovery Grants (No. RGPIN-2019-04575).

## References

[1] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *CVPR*, pages 1597–1604, 2009. 6

[2] Robert S Allison, Barbara J Gillam, and Elia Vecellio. Binocular depth discrimination and estimation beyond interaction space. *Journal of Vision*, 9(1):10–10, 2009. 2

[3] Dinkar N Bhat and Shree K Nayar. Stereo in the presence of specular reflection. In *ICCV*, pages 1086–1092, 1995. 2

[4] Ali Borji, Ming-Ming Cheng, Huaizu Jiang, and Jia Li. Salient object detection: A benchmark. *IEEE Transactions on Image Processing*, 24(12):5706–5722, 2015. 2, 6

[5] Chenglizhao Chen, Jipeng Wei, Chong Peng, and Hong Qin. Depth-quality-aware salient object detection. *IEEE Transactions on Image Processing*, 30:2350–2363, 2021. 2

[6] Chenglizhao Chen, Jipeng Wei, Chong Peng, Weizhong Zhang, and Hong Qin. Improved saliency detection in RGB-D images using two-phase depth estimation and selective deep fusion. *IEEE Transactions on Image Processing*, 29:4296–4307, 2020. 2

[7] Hao Chen, Yongjian Deng, Youfu Li, Tzu-Yi Hung, and Guosheng Lin. RGBD salient object detection via disentangled cross-modal fusion. *IEEE Transactions on Image Processing*, 29:8407–8416, 2020. 1

[8] Hao Chen and Youfu Li. Progressively complementarity-aware fusion network for RGB-D salient object detection. In *CVPR*, pages 3051–3060, 2018. 2, 6

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

[10] Hao Chen, Youfu Li, and Dan Su. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. *Pattern Recognition*, 86:376–385, 2019. 1, 6

[11] Shuhan Chen and Yun Fu. Progressively guided alternate refinement network for RGB-D salient object detection. In *ECCV*, 2020. 6

[12] Zuyao Chen, Runmin Cong, Qianqian Xu, and Qingming Huang. Depth potentiality-aware gated attention network for RGB-D salient object detection. *IEEE Transactions on Image Processing*, 2020. 1

[13] Yupeng Cheng, Huazhu Fu, Xingxing Wei, Jiangjian Xiao, and Xiaochun Cao. Depth enhanced saliency detection method. In *Proceedings of International Conference on Internet Multimedia Computing and Service*, pages 23–27, 2014. 6, 8

[14] Runmin Cong, Jianjun Lei, Huazhu Fu, Ming-Ming Cheng, Weisi Lin, and Qingming Huang. Review of visual saliency detection with comprehensive information. *IEEE Transactions on circuits and Systems for Video Technology*, 29(10):2941–2959, 2018. 1

[15] Runmin Cong, Jianjun Lei, Huazhu Fu, Junhui Hou, Qingming Huang, and Sam Kwong. Going from RGB to RGBD saliency: A depth-guided transformation model. *IEEE Transactions on Systems, Man, and Cybernetics*, 50(8):3627–3639, 2020. 2

[16] Runmin Cong, Jianjun Lei, Huazhu Fu, Qingming Huang, Xiaochun Cao, and Chunping Hou. Co-saliency detection for RGBD images based on multi-constraint feature matching and cross label propagation. *IEEE Transactions on Image Processing*, 27(2):568–579, 2018. 2

[17] Runmin Cong, Jianjun Lei, Huazhu Fu, Qingming Huang, Xiaochun Cao, and Nam Ling. HSCS: Hierarchical sparsity based co-saliency detection for RGBD images. *IEEE Transactions on Multimedia*, 21(7):1660–1671, 2018. 2

[18] Runmin Cong, Jianjun Lei, Huazhu Fu, Weisi Lin, Qingming Huang, Xiaochun Cao, and Chunping Hou. An iterative co-saliency framework for RGBD images. *IEEE Transactions on Cybernetics*, 49(1):233–246, 2019. 2

[19] Runmin Cong, Jianjun Lei, Huazhu Fu, Fatih Porikli, Qingming Huang, and Chunping Hou. Video saliency detection via sparsity-based reconstruction and propagation. *IEEE Transactions on Image Processing*, 28(10):4819–4931, 2019. 1

[20] Runmin Cong, Jianjun Lei, Changqing Zhang, Qingming Huang, Xiaochun Cao, and Chunping Hou. Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. *IEEE Signal Processing Letters*, 23(6):819–823, 2016. 6

[21] Karthik Desingh, K Madhava Krishna, Deepu Rajan, and CV Jawahar. Depth really matters: Improving visual salient region detection with depth. In *BMVC*, 2013. 2

[22] Deng-Ping Fan, Ming-Ming Cheng, Jiang-Jiang Liu, Shang-Hua Gao, Qibin Hou, and Ali Borji. Salient objects in clutter: Bringing salient object detection to the foreground. In *ECCV*, pages 186–202, 2018. 2

[23] Deng-Ping Fan, Cheng Gong, Yang Cao, Bo Ren, Ming-Ming Cheng, and Ali Borji. Enhanced-alignment measure for binary foreground map evaluation. In *IJCAI*, pages 698–704, 2018. 6

[24] Deng-Ping Fan, Zheng Lin, Zhao Zhang, Menglong Zhu, and Ming-Ming Cheng. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. *IEEE Transactions on Neural Networks and Learning Systems*, 2020. 1, 2, 6, 7, 8

[25] Deng-Ping Fan, Wenguan Wang, Ming-Ming Cheng, and Jianbing Shen. Shifting more attention to video salient object detection. In *CVPR*, pages 8554–8564, 2019. 1

[26] Deng-Ping Fan, Yingjie Zhai, Ali Borji, Jufeng Yang, and Ling Shao. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In *ECCV*, 2020. 2, 6

[27] David Feng, Nick Barnes, Shaodi You, and Chris McCarthy. Local background enclosure for RGB-D salient object detection. In *CVPR*, pages 2343–2350, 2016. 2

[28] Keren Fu, Deng-Ping Fan, Ge-Peng Ji, and Qijun Zhao. JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In *CVPR*, pages 3052–3062, 2020. 2, 6

[29] Silvio Giancola, Matteo Valenti, and Remo Sala. *A Survey on 3D Cameras: Metrological Comparison of Time-of-Flight, Structured-Light and Active Stereoscopy Technologies*. Springer, 2018. 1

[30] Chunle Guo, Chongyi Li, Jichang Guo, Chen Change Loy, Junhui Hou, Sam Kwong, and Runmin Cong. Zero-reference

deep curve estimation for low-light image enhancement. In *CVPR*, pages 1780–1789, 2020. 1

[31] Junwei Han, Hao Chen, Nian Liu, Chenggang Yan, and Xuelong Li. CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. *IEEE Transactions on Cybernetics*, 48(11):3171–3183, 2017. 6

[32] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, pages 770–778, 2016. 4, 6

[33] Qibin Hou, Ming-Ming Cheng, Xiaowei Hu, Ali Borji, Zhuowen Tu, and Philip HS Torr. Deeply supervised salient object detection with short connections. In *CVPR*, pages 3203–3212, 2017. 6

[34] Wei Ji, Jingjing Li, Miao Zhang, Yongri Piao, and Huchuan Lu. Accurate RGB-D salient object detection via collaborative learning. In *ECCV*, 2020. 1, 4, 6, 7, 8

[35] Yao Jiang, Tao Zhou, Ge-Peng Ji, Keren Fu, Qijun Zhao, and Deng-Ping Fan. Light field salient object detection: A review and benchmark. *arXiv preprint arXiv:2010.04968*, 2020. 2

[36] Ran Ju, Ling Ge, Wenjing Geng, Tongwei Ren, and Gangshan Wu. Depth saliency based on anisotropic center-surround difference. In *ICIP*, pages 1115–1119, 2014. 6, 8

[37] Ayoung Kim and Ryan M Eustice. Real-time visual SLAM for autonomous underwater hull inspection using visual saliency. *IEEE Transactions on Robotics*, 29(3):719–733, 2013. 1

[38] ByoungChul Ko, Soo Yeong Kwak, and Hyeran Byun. SVM-based salient region(s) extraction method for image retrieval. In *ICPR*, volume 2, pages 977–980, 2004. 1

[39] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. ImageNet classification with deep convolutional neural networks. In *NeurIPS*, pages 1097–1105, 2012. 6

[40] Chongyi Li, Runmin Cong, Junhui Hou, Sanyi Zhang, Yue Qian, and Sam Kwong. Nested network with two-stream pyramid for salient object detection in optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing*, 57(11):9156–9166, 2019. 2

[41] Chongyi Li, Runmin Cong, Sam Kwong, Junhui Hou, Huazhu F, Guopu Zhu, Dingwen Zhang, and Qingming Huang. ASIF-Net: Attention steered interweave fusion network for RGB-D salient object detection. *IEEE Transactions on Cybernetics*, 51(1):88–100, 2020. 2

[42] Chongyi Li, Runmin Cong, Yongri Piao, Qianqian Xu, and Chen Change Loy. RGB-D salient object detection with cross-modality modulation and selection. In *ECCV*, 2020. 1

[43] Chongyi Li, Huazhu Fu, Runmin Cong, Zechao Li, and Qianqian Xu. Nui-go: Recursive non-local encoder-decoder network for retinal image non-uniform illumination removal. In *ACMM*, pages 1478–1487, 2020. 2

[44] Gongyang Li, Zhi Liu, Linwei Ye, Yang Wang, and Haibin Ling. Cross-modal weighting network for RGB-D salient object detection. In *ECCV*, 2020. 1, 2, 6

[45] Nianyi Li, Jinwei Ye, Yu Ji, Haibin Ling, and Jingyi Yu. Saliency detection on light field. In *CVPR*, pages 2806–2813, 2014. 6

[46] Jiang-Jiang Liu, Qibin Hou, Ming-Ming Cheng, Jiashi Feng, and Jianmin Jiang. A simple pooling-based design for real-time salient object detection. In *CVPR*, pages 3917–3926, 2019. 2

[47] Nian Liu, Ni Zhang, and Junwei Han. Learning selective self-mutual attention for RGB-D saliency detection. In *CVPR*, pages 13756–13765, 2020. 6

[48] Ao Luo, Xin Li, Fan Yang, Zhicheng Jiao, Hong Cheng, and Siwei Lyu. Cascade graph neural networks for RGB-D salient object detection. In *ECCV*, 2020. 1

[49] Ran Margolin, Lihi Zelnik-Manor, and Ayellet Tal. How to evaluate foreground maps? In *CVPR*, pages 248–255, 2014. 6

[50] Yuzhen Niu, Yujie Geng, Xueqing Li, and Feng Liu. Leveraging stereopsis for saliency analysis. In *CVPR*, pages 454–461, 2012. 6, 8

[51] Youwei Pang, Lihe Zhang, Xiaoqi Zhao, and Huchuan Lu. Hierarchical dynamic filtering network for RGB-D salient object detection. In *ECCV*, 2020. 6

[52] Robert Patterson, Linda Moe, and Tiger Hewitt. Factors that affect depth perception in stereoscopic displays. *Human Factors*, 34(6):655–667, 1992. 2

[53] Houwen Peng, Bing Li, Weihua Xiong, Weiming Hu, and Rongrong Ji. RGBD salient object detection: a benchmark and algorithms. In *ECCV*, pages 92–109, 2014. 2, 6, 8

[54] Yongri Piao, Wei Ji, Jingjing Li, Miao Zhang, and Huchuan Lu. Depth-induced multi-scale recurrent attention network for saliency detection. In *ICCV*, pages 7254–7263, 2019. 1, 2, 6, 7, 8

[55] Yongri Piao, Zhengkun Rong, Miao Zhang, Weisong Ren, and Huchuan Lu. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In *CVPR*, pages 9060–9069, 2020. 6

[56] Xuebin Qin, Zichen Zhang, Chenyang Huang, Chao Gao, Masood Dehghan, and Martin Jagersand. BASNet: Boundary-aware salient object detection. In *CVPR*, pages 7479–7489, 2019. 2

[57] Liangqiong Qu, Shengfeng He, Jiawei Zhang, Jiandong Tian, Yandong Tang, and Qingxiong Yang. RGBD salient object detection via deep fusion. *IEEE Transactions on Image Processing*, 26(5):2274–2285, 2017. 2, 6

[58] Jianqiang Ren, Xiaojin Gong, Lu Yu, Wenhui Zhou, and Michael Ying Yang. Exploiting global priors for RGB-D saliency detection. In *CVPR Workshops*, pages 25–32, 2015. 2

[59] Zhixiang Ren, Shenghua Gao, Liang-Tien Chia, and Ivor Wai-Hung Tsang. Region-based saliency detection and its application in object recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(5):769–779, 2013. 1

[60] Florian Schroff, Dmitry Kalenichenko, and James Philbin. FaceNet: A unified embedding for face recognition and clustering. In *CVPR*, pages 815–823, 2015. 5

[61] Ling Shao and Michael Brady. Specific object retrieval based on salient regions. *Pattern Recognition*, 39(10):1932–1948, 2006. 1

[62] Riku Shigematsu, David Feng, Shaodi You, and Nick Barnes. Learning RGB-D salient object detection using background enclosure, depth contrast, and top-down features. In *ICCV Workshops*, pages 2749–2757, 2017. 2

[63] Ting-Chun Wang, Alexei A Efros, and Ravi Ramamoorthi. Occlusion-aware depth estimation using light-field cameras. In *ICCV*, pages 3487–3495, 2015. 1

[64] W Williem and In Kyu Park. Robust light field depth estimation for noisy scene with occlusion. In *CVPR*, pages 4396–4404, 2016. 2

[65] Sanghyun Woo, Jongchan Park, Joon-Young Lee, and In So Kweon. CBAM: Convolutional block attention module. In *ECCV*, pages 3–19, 2018. 5

[66] Zhe Wu, Li Su, and Qingming Huang. Cascaded partial decoder for fast and accurate salient object detection. In *CVPR*, pages 3907–3916, 2019. 3, 6

[67] Zhe Wu, Li Su, and Qingming Huang. Stacked cross refinement network for edge-aware salient object detection. In *ICCV*, pages 7264–7273, 2019. 2

[68] Matthew D Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *ECCV*, pages 818–833, 2014. 5

[69] Jing Zhang, Deng-Ping Fan, Yuchao Dai, Saeed Anwar, Fatemeh Sadat Saleh, Tong Zhang, and Nick Barnes. UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders. In *CVPR*, pages 8582–8591, 2020. 2, 6

[70] Miao Zhang, Sun Xiao Fei, Jie Liu, Shuang Xu, Yongri Piao, and Huchuan Lu. Asymmetric two-stream architecture for accurate RGB-D saliency detection. In *ECCV*, 2020. 6

[71] Miao Zhang, Wei Ji, Yongri Piao, Jingjing Li, Yu Zhang, Shuang Xu, and Huchuan Lu. LFNet: Light field fusion network for salient object detection. *IEEE Transactions on Image Processing*, 29:6276–6287, 2020. 1

[72] Miao Zhang, Jingjing Li, Wei Ji, Yongri Piao, and Huchuan Lu. Memory-oriented decoder for light field salient object detection. In *NeurIPS*, pages 896–906, 2019. 1

[73] Miao Zhang, Weisong Ren, Yongri Piao, Zhengkun Rong, and Huchuan Lu. Select, supplement and focus for RGB-D saliency detection. In *CVPR*, pages 3472–3481, 2020. 6

[74] Miao Zhang, Yu Zhang, Yongri Piao, Beiqi Hu, and Huchuan Lu. Feature reintegration over differential treatment: A top-down and adaptive fusion network for RGB-D salient object detection. In *ACMM*, 2020. 6

[75] Qijian Zhang, Runmin Cong, Junhui Hou, Chongyi Li, and Yao Zhao. CoADNet: Collaborative aggregation-and-distribution networks for co-salient object detection. In *NeurIPS*, 2020. 2

[76] Zhengyou Zhang. Microsoft Kinect sensor and its effect. *IEEE Multimedia*, 19(2):4–10, 2012. 1, 2

[77] Jiawei Zhao, Yifan Zhao, Jia Li, and Xiaowu Chen. Is depth really necessary for salient object detection? In *ACMM*, 2020. 2

[78] Jia-Xing Zhao, Yang Cao, Deng-Ping Fan, Ming-Ming Cheng, Xuan-Yi Li, and Le Zhang. Contrast prior and fluid pyramid integration for RGBD salient object detection. In *CVPR*, pages 3927–3936, 2019. 2, 6

[79] Jia-Xing Zhao, Jiang-Jiang Liu, Deng-Ping Fan, Yang Cao, Jufeng Yang, and Ming-Ming Cheng. EGNet: Edge guidance network for salient object detection. In *ICCV*, pages 8779–8788, 2019. 2

[80] Shanshan Zhao, Huan Fu, Mingming Gong, and Dacheng Tao. Geometry-aware symmetric domain adaptation for monocular depth estimation. In *CVPR*, pages 9788–9798, 2019. 6

[81] Xiaoqi Zhao, Lihe Zhang, Youwei Pang, Huchuan Lu, and Lei Zhang. A single stream network for robust and real-time RGB-D salient object detection. In *ECCV*, 2020. 6

[82] Tao Zhou, Deng-Ping Fan, Ming-Ming Cheng, Jianbing Shen, and Ling Shao. RGB-D salient object detection: A survey. *Computational Visual Media*, pages 1–33, 2021. 2

[83] Chunbiao Zhu, Xing Cai, Kan Huang, Thomas H Li, and Ge Li. PDNet: Prior-model guided depth-enhanced network for salient object detection. In *ICME*, pages 199–204, 2019. 6

[84] Chunbiao Zhu, Ge Li, Xiaoqiang Guo, Wenmin Wang, and Ronggang Wang. A multilayer backpropagation saliency detection algorithm based on depth mining. In *CAIP*, pages 14–23, 2017. 6

[85] Chunbiao Zhu, Ge Li, Wenmin Wang, and Ronggang Wang. An innovative salient object detection using center-dark channel prior. In *ICCV Workshops*, pages 1509–1515, 2017. 6