



# CFIDNet: cascaded feature interaction decoder for RGB-D salient object detection

Tianyou Chen<sup>1</sup> · Xiaoguang Hu<sup>1</sup> · Jin Xiao<sup>1</sup> · Guofeng Zhang<sup>1</sup> · Shaojie Wang<sup>1</sup>

Received: 24 May 2021 / Accepted: 12 December 2021 / Published online: 7 January 2022  
© The Author(s), under exclusive licence to Springer-Verlag London Ltd., part of Springer Nature 2022

## Abstract

Compared with RGB salient object detection (SOD) methods, RGB-D SOD models show better performance in many challenging scenarios by leveraging spatial information embedded in depth maps. However, existing RGB-D SOD models prone to ignore the modality-specific characteristics and fuse multi-modality features by simple element-wise addition or multiplication. Thus, they may induce noise-degraded saliency maps when encountering inaccurate or blurred depth images. Besides, many models adopt the U-shape architecture to integrate multi-level features layer-by-layer. Despite the fact that low-level features can be gradually polished, little attention has been paid to enhance high-level features, which may lead to suboptimal results. In this paper, we propose a novel network named CFIDNet to tackle the above problems. Specifically, we design the feature-enhanced module to excavate informative depth cues from depth images and enhance the RGB features by employing complementary information between RGB and depth modalities. Besides, we propose the feature refinement module to exploit multi-scale complementary information between multi-level features and polish these features by applying residual connections. The cascaded feature interaction decoder (CFID) is then proposed to refine multi-level features iteratively. Equipped with these proposed modules, our CFIDNet is capable of segmenting salient objects accurately. Experimental results on 7 widely used benchmark datasets validate that our CFIDNet achieves highly competitive performance over 15 state-of-the-art models in terms of 8 evaluation metrics. Our source code will be publicly available at <https://github.com/clelouch/CFIDNet>.

**Keywords** Salient object detection · RGB-D images · Cross-modality feature fusion · Cascaded refinement

## 1 Introduction

Salient object detection (SOD) aims to locate and segment the most visually prominent regions in an image [1, 2]. As a widely used preprocessing technique, SOD plays an important role in numerous computer vision tasks, such as object recognition [3], image editing [4], visual tracking [5–7], person reidentification [8], video analysis [9, 10], and thumbnail creation [11]. Traditional approaches [12–17] mainly rely on handcrafted features to excavate low-level details for saliency prediction but cannot capture high-level semantic knowledge. Thus, these methods are not able to locate salient objects accurately and segment them precisely when faced with complex scenarios (e.g., cluttered backgrounds).

Recently, owing to the outstanding representation ability of convolutional neural networks (CNNs), various CNN-based RGB SOD methods [18–34] have been proposed.

---

✉ Jin Xiao  
xiaojin@buaa.edu.cn

Tianyou Chen  
chentianyou@buaa.edu.cn

Xiaoguang Hu  
xiaoguang@buaa.edu.cn

Guofeng Zhang  
gfbzhang@buaa.edu.cn

Shaojie Wang  
sjwang@buaa.edu.cn

<sup>1</sup> State Key Laboratory of Virtual Reality Technology and Systems, School of Automation Science and Electrical Engineering, Beihang University, Beijing 100191, China

Benefiting from the hierarchical network architecture, these methods show promising performance in capturing multi-level features and outperform traditional counterparts by a significant margin. However, as pointed out in [35, 36], the performance of RGB SOD methods may dramatically decrease when encountering some challenging scenarios (e.g., transparent regions, cluttered backgrounds, low contrast). To solve this problem, depth images are introduced to provide geometrical information, which has been proven to be an effective approach in improving the performance of SOD. In the past few years, various models [35–49] have been proposed to boost the performance of SOD by leveraging both RGB and depth information.

Despite the superior performance of these methods, there are two main issues remaining unsolved. First, existing RGB-D SOD methods [35, 39, 43, 50] mainly attach equal importance to RGB features and depth ones and integrate these multi-modality features by simple element-wise addition, multiplication or concatenation. However, as shown in Fig. 1, depth maps may be easily influenced by environment and full of noises, which can be attributed to multiple factors during the acquisition of depth images [37, 38], e.g., high sensor temperature, unstable devices, bright background illumination and reflectivity of the observed objects. As a result, these noise-degraded depth images cannot provide useful spatial cues to improve the performance of SOD and may even mislead the saliency prediction. Second, as pointed out in many previous methods [20, 24], due to the hierarchical architecture of CNN and sub-sampling operations (e.g., pooling and convolution) applied in networks, low-level features have larger spatial size and maintain affluent spatial details, which are helpful to sharpen the boundaries of salient objects. Compared with low-level features, high-level ones are coarser in boundaries but contain more semantic knowledge, which is beneficial to locate the salient objects

and suppress background noises embedded in low-level features. Based on this observation, many models [18, 23, 24, 27] adopt the U-shape architecture, where high-level semantic knowledge is gradually transmitted to shallower stages to better locate salient regions. However, little attention has been paid to polish high-level features. Consequently, the continuous accumulation of low-quality (e.g., inaccurate or blurred) deeper features may result in performance degradation.

To tackle the above issues, we propose a novel network named cascaded feature interaction decoder (CFIDNet) for RGB-D SOD.

First, we argue that the depth features are less important and it is more favorable to use them as informative aid to assist SOD because depth images may be of low quality and hence pose a risk to accurate saliency prediction. Besides, RGB features generally maintain more semantic information. For example, a green object is much more likely to be a plant than a red one. Thus, to better integrate multi-modality (i.e., RGB and depth) features, we propose a Feature-enhanced Module (FEM). In the FEM, we employ the attention mechanism to exploit informative cues from depth features and integrate the cross-modality features by concatenation. A convolutional block is then applied to excavate complementary information. Afterward, a residual connection is adopted to enhance RGB features by leveraging the complementary features. The proposed FEM, though simple, is effective in generating robust fused features.

Second, we propose the feature refinement module (FRM) to simultaneously refine multi-level features. The FRM is partly inspired by DANet [50], where a pyramidally attended feature extraction (PAFE) module is placed on the top of the backbone to capture multi-scale context for different objects. However, as pointed out in MINet [32], each convolutional layer can only capture a specific scale contextual information. Thus, only exploiting multi-scale information from the highest feature may be insufficient and lead to suboptimal results. Besides, as pointed out in S2MA [74], typical CNN-based models can hardly model the long-range dependency. In this paper, we tackle the above issues in a “rescaling-integrating-refining-strengthening” manner [87] and focus on the “refining” process. Concretely, we divide the input multi-level features into three groups and use the FRM to polish them group by group. In the proposed FRM, we integrate multi-level features by concatenation and employ a non-local block [52] to capture long-range dependencies. Afterward, we utilize a tiny U-shape block to excavate intra-layer multi-scale information, which is then integrated with the input multi-level features of FRM to generate features robust to scale variation. In this way, we can effectively refine multi-level features.

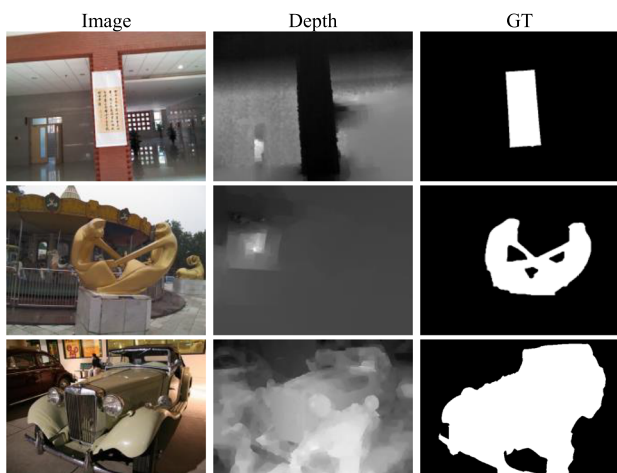


Fig. 1 Several representative cases with misleading depth maps

Third, we develop the cascaded feature interaction decoder (CFID) to refine multi-level features iteratively. The proposed CFID contains multiple sub-decoders, each of which uses three FRMs to progressively polish features of all levels. Equipped with the above-mentioned modules, our proposed CFIDNet is able to accurately locate and segment salient objects.

To demonstrate the superior performance of the proposed CFIDNet, we conduct extensive experiments on seven widely used RGB-D saliency detection benchmark datasets. Besides, visualizations of feature maps are also provided to prove the effectiveness of the proposed modules more intuitively. Those experimental results validate that our CFIDNet is effective in generating high-quality saliency maps. In summary, our paper makes four contributions:

- We design the feature-enhanced module (FEM) to integrate multi-modality features. The proposed module is able to excavate complementary information between RGB and depth modality and enhance RGB features to generate robust fused ones.
- We propose the feature refinement module (FRM) to correct and polish multi-level features and exploit multi-scale information to make them competent to deal with scale variation of different salient objects.
- We develop the cascaded feature interaction decoder (CFID) to iteratively refine multi-level features. The CFID is composed of multiple feature interaction decoders (FIDs), each of which employs three FRMs to progressively integrate multi-level features and refine them. The repeatedly applied FIDs not only suppress the background noises for low-level features but also sharpen the boundaries for high-level features.
- Extensive experimental results on seven commonly used RGB-D saliency detection datasets demonstrate that the CFIDNet achieves highly competitive performance over 15 state-of-the-art approaches under 8 standard evaluation metrics, which validates the effectiveness and superiority of our proposed model.

## 2 Related work

### 2.1 Salient object detection

Over the past few years, SOD has attracted wide interest due to its outstanding performance in various computer vision tasks [1–11]. Traditional SOD algorithms generally solve this problem by exploring handcrafted features, e.g., foreground consistency [15], center prior [14, 17], histograms [16, 53] and so on. However, these heuristic saliency information and low-level handcrafted features

cannot capture semantic knowledge, which makes these algorithms not able to generate accurate saliency maps, especially when encountering challenging scenarios. Recently, various deep learning-based models have been proposed to solve the problem. Benefiting from the powerful CNNs, these models outperform traditional counterparts by a large margin.

Zhang et al. (Amulet) [54] generate saliency maps by integrating multi-level features into various resolutions to simultaneously incorporate global semantics and local details. Want et al. (SRM) [55] employ the pyramid pooling module [56] to obtain features with richer global context information and predict saliency maps by using multistage refinement mechanism. Zhang et al. (UCF) [22] develop a novel hybrid upsampling method and a reformulated dropout to generate accurate saliency prediction. Luo et al. (NLDF) [21] develop a novel network to aggregate global and local information through a grid structure. Hou et al. (DSS) [20] introduce short connections to the skip-layer structure to make better use of features extracted from CNNs. Zhang et al. (BMPM) [25] develop a bi-directional structure to pass message between multi-level features. Liu et al. (PiCANet) [57] recurrently excavate global and local contextual attention and generate saliency prediction by incorporating it with an encoder-decoder architecture. Feng et al. [58] implement a global perception module to roughly capture salient objects. An attentive feedback module is then introduced to refine the coarse detection scale-by-scale. Qin et al. (BASNet) [27] build a predict-refine model by stacking two U-shape networks sequentially. A hybrid loss function composed of a standard binary cross-entropy loss, a structural similarity loss and an IoU (Intersection-over-Union) loss is also designed to obtain accurate saliency map with sharper boundaries. Wu et al. (CPD) [28] achieve fast and accurate saliency prediction by proposing a cascaded optimization mechanism, where initial saliency map generated by the first branch is utilized to refine features of the second branch. Zhao et al. (EGNet) [23] explore the complementarity between salient contour information and salient object cues. Zhao et al. (PFAN) [59] adopt channel-wise attention and spatial attention modules to focus on the most informative parts of features and design the context-aware pyramid feature extraction module to capture rich context information. Liu et al. (PoolNet) [24] adopt the encoder-decoder architecture and develop a multi-scale feature aggregation module to better fuse multi-level features. Wu et al. [60] explore the interrelations of edge and segmentation and stack multiple cross-refinement units to simultaneously refine multi-level features of contour detection and saliency segmentation. Liu et al. (DFI) [61] show the similarities shared by SOD, edge detection and skeleton extraction and develop a novel network to solve the three

tasks jointly. Wei et al. ( $F^3$ Net) [29] propose a cascaded feedback decoder to aggregate multi-level features gradually and refine them iteratively. Gao et al. (CSNet) [30] propose the generalized OctConv (gOctConv) and build an extremely light-weighted network utilizing the proposed gOctConv. Zhou et al. (ITSD) [31] propose a two-stream decoder to explore multiple cues of the contour and saliency maps. Qin et al. ( $U^2$  - Net) develop a two-level nested U-shape network, which is able to be trained from scratch and shows comparable or even better performance than models with pretrained backbones.

## 2.2 RGB-D salient object detection

Previous traditional methods for RGB-D SOD mainly rely on handcrafted features [62, 63] extracted from RGB and depth images. Basically, these algorithms exploit contrast-based cues (e.g., color, edge, and texture) to calculate the saliency confidence of a local region. For instance, Cheng et al. (DES) [62] conduct pixel clustering and measure each cluster's saliency confidence using three cues (i.e., depth contrast, color contrast and spatial bias). Thus, the final saliency prediction can be generated by combining these cues. Peng et al. [64] propose a multistage RGB-D saliency estimation method to generate saliency prediction by combining depth cues and appearance information. Song et al. [65] develop a multi-scale discriminative saliency fusion method for RGB-D SOD. Feng et al. [66] design the local background enclosure feature to directly measure salient structure from depth. Ju et al. [67] use the anisotropic center-surround difference to define depth saliency confidence of a point and utilize the depth and center priors for further refinement.

Recently, various deep learning-based models have been proposed. Compared with previous methods, these models show better performance, hence becoming a mainstream trend in RGB-D SOD. Hussain et al. [86] propose a novel architecture employing densely deformable convolutions to capture the salient objects' regions. Zhu et al. (PDNet) [68] develop a depth-enhanced model consisting of a master network and a sub-network, where the sub-network is used to process depth map and enhance the robustness of the master network. Wang et al. (AFNet) [39] propose a two-streamed model to generate a saliency map from each modality separately and develop a saliency fusion module to learn a switch map to fuse the generated saliency maps. Zhao et al. (CPFP) [41] propose a contrast loss to leverage the contrast prior for depth map enhancement. Besides, a fluid pyramid integration strategy is proposed to utilize multi-scale cross-modal features for saliency prediction. Piao et al. (DMRA) [43] combine depth information with multi-scale cues for saliency prediction and boost the

performance by utilizing a recurrent attention module. Chen et al. (MMCI) [49] develop a multi-path multi-modal fusion network. Zhang et al. (FRDT) [44] propose a top-down multi-level fusion structure. Piao et al. (A2dele) [38] develop a depth distiller to transfer depth knowledge from depth features to RGB features. Zhang et al. (ATSA) [40] consider the inherent differences between depth and RGB modality and develop an asymmetric two-stream architecture to predict saliency map accurately. Zhai et al. (BBSNet) [35] design a bifurcated backbone strategy network to exploit multi-level features in a cascaded refinement manner and suppress noises in shallower layers. Chen et al. (DPANet) [36] design a novel network to address unreliable depth images. Jin et al. (CDNet) [37] design a network robust to the unstable quality of depth maps. Ji et al. (CoNet) [69] propose a collaborative learning framework, where saliency, edge and depth are utilized for saliency detection. Fan et al. ( $D^3$ Net) [70] design a network to learn to automatically discard depth maps of low quality. Zhao et al. (DANet) [50] build a single stream network with depth-enhanced dual attention module to achieve real-time and robust saliency prediction. Fu et al. (JL-DCF) [48] use a Siamese network to learn from both RGB and depth modalities. Pang et al. (HDFNet) [45] propose a hierarchical dynamic filtering network for RGB-D SOD. A hybrid enhanced loss is then leveraged to effectively improve the detection performance. Li et al. (HAINet) [51] adopt the encoder-decoder architecture and leverage the hierarchical alternate interaction module to effectively highlight salient objects and mitigate distractors in depth images.

## 3 Proposed method

### 3.1 Overview of the proposed CFIDNet

The overall pipeline of the proposed CFIDNet is illustrated in Fig. 2. The CFIDNet consists of two kinds of modules including the CFID and the backbone encoder. Following many previous methods [18, 23, 24, 61], we employ the commonly used ResNet-50 [71] as the backbone network and discard the average pooling layer and fully connected layer. Besides, the dilation rates of  $3 \times 3$  convolutional layers in the last residual block are set to 2 to obtain feature maps with larger resolution. For simplicity, the depth map is converted to a three-channel image by replicating the single channel image into three channels.

Given a pair of input images with a spatial resolution of  $H \times W$ , we extract features at five stages from RGB and depth modalities, respectively. Since the extracted side-output features have different resolutions and channel

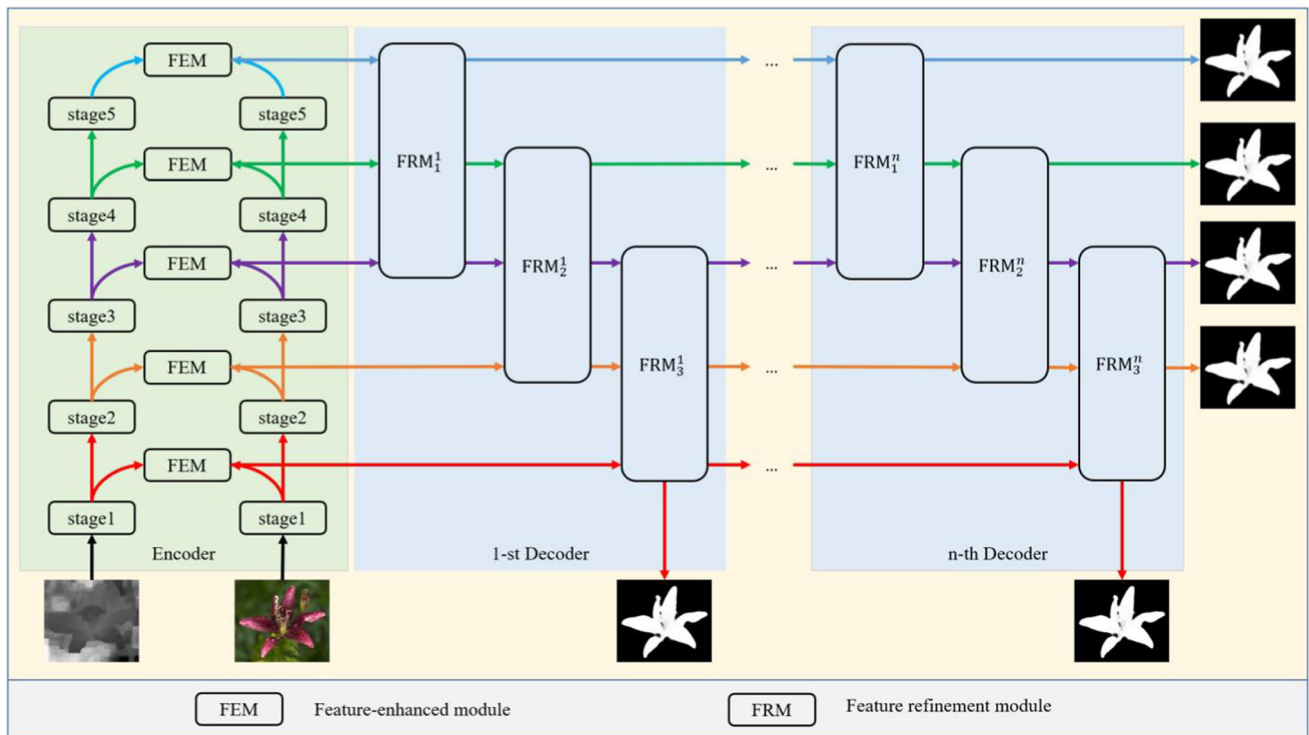


Fig. 2 The overall pipeline of our proposed CFIDNet

numbers, we use  $1 \times 1$  convolutional layers to reduce the channel number to 64, which not only facilitates various element-wise operations but also is effective in reducing computation and memory overhead. It is worth noting that the  $1 \times 1$  convolutional layers are omitted in Fig. 2 for conciseness. The extracted side-output features from RGB and depth modalities are then denoted as  $\{F_i^r | i = 1, 2, 3, 4, 5\}$  and  $\{F_i^d | i = 1, 2, 3, 4, 5\}$ , respectively. The resolutions of these features can be obtained by computing:

$$\begin{cases} \frac{H}{2^i} \times \frac{W}{2^i}, & i = 1, 2, 3, 4 \\ \frac{H}{2^{i-1}} \times \frac{W}{2^{i-1}}, & i = 5 \end{cases} \quad (1)$$

Each pair of multi-modality features (i.e.,  $F_i^r$  and  $F_i^d$ ) is then fed into a FEM to generate the corresponding feature enhanced feature, which is denoted as  $F_i^0 (i = 1, 2, 3, 4, 5)$ . In the FEM, the complementary information between the multi-modality features is excavated to generate robust fused feature. After obtaining these multi-level fused features, a sequence of FIDs is leveraged to refine them iteratively. It is worth noting that the spatial resolution of an input feature of an FID is the same as that of the corresponding output one. Thus, the outputs of the former FID can be directly utilized as the inputs of the latter one. The CFIDNet can be trained in an end-to-end manner without using any preprocessing (e.g., HHA [72]) or postprocessing (e.g., CRF [73]) methods. Besides, inspired by [18, 20, 34],

aside from the dominant losses corresponding to side-output features with the largest spatial resolution, other side-output features of the last decoder are also utilized to compute the auxiliary losses to facilitate optimization. Experimental results show that equipped with two FIDs, the CFIDNet is able to segment salient objects accurately and achieves an average inference speed of 22FPS on a single NVIDIA Titan Xp GPU.

### 3.2 Feature refinement module (FRM)

The structure of FRM is illustrated in Fig. 3. In the proposed FRM, we first integrate multi-level features from three adjacent layers by concatenation. Considering the adjacent resolution, the fusion strategy is effective in avoiding interference caused by large resolution differences. The input high-level, middle-level and low-level features are denoted as  $F_h, F_m, F_l$ , respectively. Since the spatial resolution of these features is different, we resize  $F_h$  to the same size as  $F_m$  by using bilinear interpolation operation. A convolutional layer is then applied for refinement. Similarly, a convolutional layer with a stride of 2 pixels is applied to  $F_l$ . The fusion process of these three features is formulated as:

$$f_h = C_3(Up(F_h, F_m)), f_l = C_3^s(F_l), \quad (2)$$

$$f_c = cat(f_h, F_m, f_l), \quad (3)$$

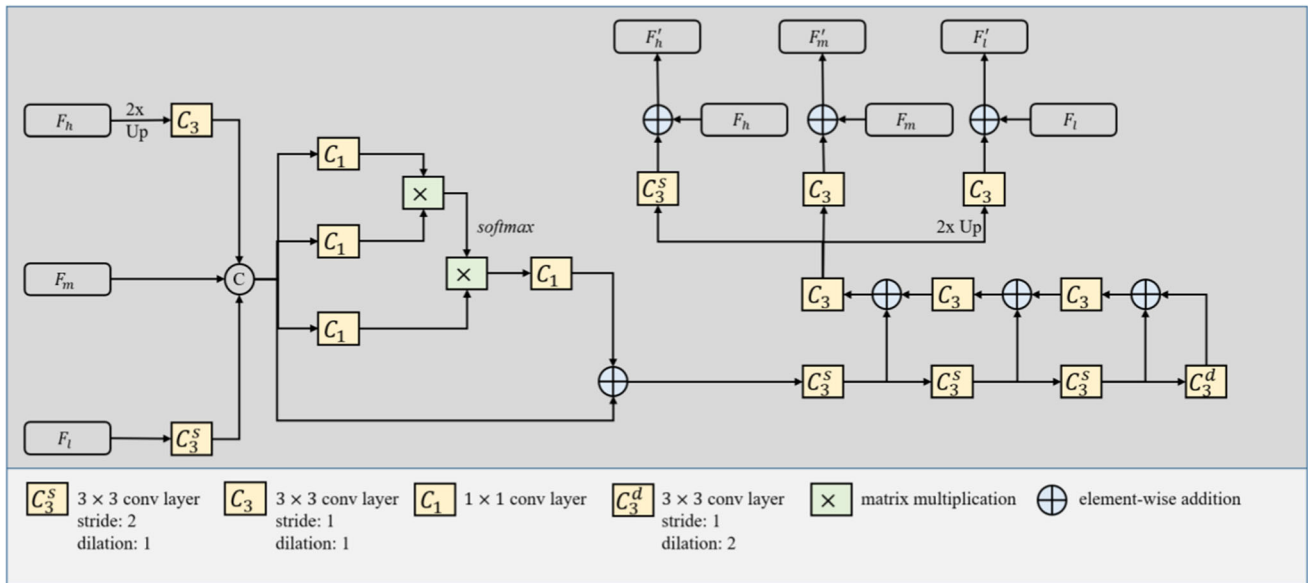


Fig. 3 Illustration of FRM

where  $C_3$  denotes a  $3 \times 3$  convolutional layer,  $Up(F_h, F_m)$  is the bilinear interpolation operation, which is used to resize the  $F_h$  to the same size as  $F_m$ ,  $C_3^s$  denotes a  $3 \times 3$  convolutional layer with 2 stride, and  $cat$  is the concatenation operation.

After obtaining the fused feature, we employ a non-local block [52] to capture long-range dependencies. Thus, long-range global context can be exploited for refinement. Let the channel, width, height of  $f_c$  be denoted by  $c, h, w$ . The process can be depicted as:

$$f_q = C_1^m(f_c), f_k = C_1^m(f_c), f_v = C_1^m(f_c), \tag{4}$$

$$f_m = f_q \boxtimes f_k^T, \tag{5}$$

$$f_r = softmax(f_m \boxtimes f_v^T), \tag{6}$$

$$f_0 = f_c + C_1(f_r), \tag{7}$$

where  $C_1$  denotes a  $1 \times 1$  convolutional layer and the input channel number is equal to the output channel number,  $C_1^m$  is a  $1 \times 1$  convolutional layer and the output channel number is set to  $1/8$  of the input channel number to reduce computation and memory overhead,  $f_q f_k, f_v$  and  $f_m \in \mathbb{R}^{c \times (hw)}$ ,  $\boxtimes$  is the matrix multiplication operation, and  $softmax$  is applied in the column of  $(f_m \boxtimes f_v^T)$ .

After obtaining the  $f_0$ , we use a tiny U-shape block to exploit multi-scale information. Different from [48] that uses an inception-like structure and [51] that uses multiple dilated convolutions with incremental dilation rates, our proposed U-shape block is not only effective in capturing multi-scale information but also beneficial to reduce computation costs, since most operations are conducted on

subsampled feature maps. The multi-scale feature extraction process can be described as:

$$f_i = \begin{cases} C_3^s(f_{i-1}), i = 1, 2, 3 \\ C_3^d(f_{i-1}), i = 4 \end{cases} \tag{8}$$

$$\bar{f}_i = \begin{cases} C_3(Up(\bar{f}_{i+1}, f_i) + f_i), i = 1, 2 \\ C_3(f_3 + f_4), i = 3 \end{cases} \tag{9}$$

where  $C_3^d$  is a  $3 \times 3$  convolutional layer with the dilation rate of 2 and stride of 1,  $\bar{f}_1$  denotes the output. The  $\bar{f}_1$  is then used to refine  $F_h, F_m, F_l$ . The refinement process can be formulated as:

$$\begin{cases} F'_h = F_h + C_3^s(\bar{f}_1), \\ F'_m = F_m + C_3(\bar{f}_1), \\ F'_l = F_l + C_3(Up(\bar{f}_1, F_l)), \end{cases} \tag{10}$$

where  $F'_h, F'_m, F'_l$  are the output features of FRM.

The proposed FRM brings us three benefits. First, we can exploit multi-scale information from each group to effectively alleviate the scale variation issue. Besides, the FRM only processes features of adjacent layers, which is effective in avoiding the interference caused by large resolution differences [32]. Second, complementary information can be excavated to enhance these multi-level features. For example, low-level details (e.g., sharp boundaries) can be transferred to deeper stages while high-level knowledge (e.g., object location) can be delivered to shallower stages. Thus, we can refine features of all levels simultaneously. While previous works mainly focus on the refinement of low-level features, the enhancement of high-level ones is proven effective in our paper. Third, we focus on the

refinement of the integrated feature, which can reduce the computation and memory overhead.

### 3.3 Cascaded feature interaction decoder (CFID)

The CFID is composed of multiple FIDs and each FID consists of three FRMs. Taking the  $i$ -th FID for example, the input features are denoted as  $F_j^{i-1} (j = 1, 2, 3, 4, 5)$ . The outputted features of this decoder can be obtained by computing:

$$\begin{cases} F_{3,1}^{i-1}, F_{4,1}^{i-1}, F_{5,1}^{i-1} = FRM_1^i(F_3^{i-1}, F_4^{i-1}, F_5^{i-1}) \\ F_{2,1}^{i-1}, F_{3,2}^{i-1}, F_{4,2}^{i-1} = FRM_2^i(F_2^{i-1}, F_{3,1}^{i-1}, F_{4,1}^{i-1}) \\ F_{1,1}^{i-1}, F_{2,2}^{i-1}, F_{3,3}^{i-1} = FRM_3^i(F_1^{i-1}, F_{2,1}^{i-1}, F_{3,2}^{i-1}) \end{cases} \quad (11)$$

$$F_1^i, F_2^i, F_3^i, F_4^i, F_5^i = F_{1,1}^{i-1}, F_{2,2}^{i-1}, F_{3,3}^{i-1}, F_{4,2}^{i-1}, F_{5,1}^{i-1} \quad (12)$$

where  $FRM_j^i$  is the  $j$ -th FRM of the  $i$ -th decoder,  $F_j^i (j = 1, 2, 3, 4, 5)$  are the output features. By using the FID, features are progressively integrated from deeper stages to shallower stages. The aggregated feature with the finest spatial resolution (i.e.,  $F_1^i$ ) is utilized to generate a saliency map for supervision. The outputted features of the  $i$ -th decoder are then directly fed to  $(i + 1)$ -th decoder for further refinement. By continuously polishing these multi-level features, the CFIDNet can refine features from all levels and is able to generate finer saliency maps.

To validate the effectiveness of the proposed CFID, we visualize feature maps at different places in CFIDNet with two FIDs. As shown in Fig. 4, we provide the visualizations of  $F_2^0, F_2^1$ , and  $F_2^2$  feature maps. It can be clearly seen from Fig. 4 that the CFID is not only effective in suppressing background noises by leveraging high-level semantics but also useful to sharpen the salient boundaries by utilizing low-level details.

### 3.4 Feature-enhanced module (FEM)

As show in Fig. 1, depth maps may be of low quality and hence pose a risk to saliency detection. Thus, to effectively integrate RGB and depth features, it is desirable to exploit

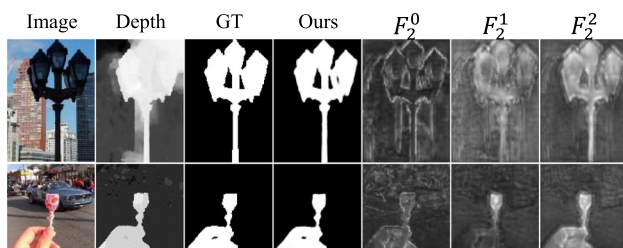


Fig. 4 Visualizations of feature maps at different places in CFIDNet with two FIDs

useful complementary information between RGB and depth modalities to enhance RGB features. To solve this problem, we propose the FEM. The structure of FEM is illustrated in Fig. 5.

The FEM is inspired by the JLDCF [48]. In JLDCF, it has been proven that an RGB SOD model can sometimes perform well in the depth view, which means the appearance information in depth images can be also leveraged to excavate the semantic knowledge. Therefore, we employ the attention mechanism to exploit informative cues from depth features. When depth images contain useful sematic information, we can effectively excavate them to enhance the original depth features. In this process, channels showing higher response to salient regions are highlighted, which implies the categories of the salient objects. Besides, when depth images are blurred, the appearance information can be hardly used to excavate informative cues, which means the channels of blurred features will not be highlighted. Therefore, we can focus on the most informative regions of depth images using the attention operations. Then, we integrate the cross-modality features by concatenation. Afterward, a residual block is applied for further refinement. The refined depth feature is integrated with RGB feature via concatenation. Two  $3 \times 3$  convolutional layers are applied to exploit complementary information between RGB and depth modalities. Thus, we can obtain the enhanced RGB feature by combining the complementary information and the original RGB feature using element-wise addition. The entire process can be formulated as:

$$\hat{f}_d = f_d \times \text{sigmoid}(fc_2(fc_1(\text{pool}(f_d)))) \quad (13)$$

$$f_d^e = \hat{f}_d + C_3(\hat{f}_d) \quad (14)$$

$$f_r^e = f_r + C_3(C_3(\text{cat}(f_r, f_d^e))) \quad (15)$$

where  $f_d$  is the depth feature,  $f_r$  is the RGB feature,  $\text{pool}$  is a global average pooling layer,  $fc_1$  and  $fc_2$  are two fully connected layers,  $\text{sigmoid}$  is the sigmoid function,  $f_r^e$  is the enhanced RGB feature.

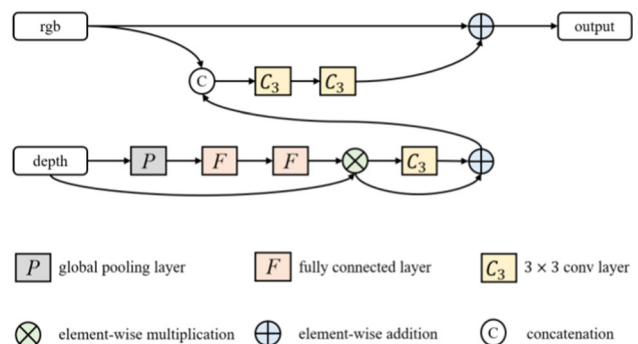


Fig. 5 Illustration of FEM

### 3.5 Loss function

Our loss function consists of two parts, i.e., the standard binary cross-entropy (BCE) loss and the IoU [27] loss. As the most commonly used loss function in SOD task, the BCE loss calculates the per-pixel loss independently without paying attention to the global structure of the image. Besides, it is not able to prevent interference caused by imbalance ratios between foreground and background regions. Thus, the IoU loss is introduced to emphasize the global structure. For an generated saliency map  $S$  with a spatial resolution  $H \times W$ , the training loss can be obtained by calculating:

$$\mathcal{L}_{BCE} = \frac{-1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W |\log(S_{i,j}) \times G_{i,j} + \log(1 - S_{i,j}) \times (1 - G_{i,j})|, \quad (16)$$

$$\mathcal{L}_{IoU} = -\frac{\sum_{i=1}^H \sum_{j=1}^W (S_{i,j} \times G_{i,j})}{\sum_{i=1}^H \sum_{j=1}^W (S_{i,j} + G_{i,j} - S_{i,j} \times G_{i,j})} + 1, \quad (17)$$

$$\mathcal{L} = \mathcal{L}_{BCE} + \mathcal{L}_{IoU}, \quad (18)$$

where  $G$  denotes the groundtruth,  $S_{i,j}$  is the saliency confidence for pixel in  $(i,j)$  of the saliency map,  $G_{i,j}$  is the corresponding mask label.

Inspired by [20, 29], for CFIDNet with  $n$  FIDs, we calculate  $n$  dominant losses and 4 auxiliary losses. Specifically, a  $1 \times 1$  convolutional layer is applied to the side-output feature with the largest spatial resolution in each FID to generate a saliency map, which corresponds to a dominant loss. Besides, the rest side-output features of the last FID are also used to facilitate optimization. Thus, the whole process is defined as:

$$S_j^i = Up\left(\text{sigmoid}\left(C_1^p\left(F_j^i\right)\right), G\right), \quad (19)$$

$$\mathcal{L}_{sum} = \sum_{i=1}^n \mathcal{L}(S_1^i, G) + \sum_{j=2}^5 \lambda_j \times \mathcal{L}(S_j^n, G), \quad (20)$$

$$\lambda_j = \frac{w_j^n}{w_1^n}, \quad (21)$$

where  $C_1^p$  is a  $1 \times 1$  convolutional layer and the output channel number is set to 1,  $\mathcal{L}_{sum}$  is the final loss, and  $w_j^i$  is the width of  $F_j^i$ .

## 4 Experiments and results

### 4.1 Implementation details

We use PyTorch toolbox to implement the CFIDNet and conduct experiments on 8 widely used RGB-D benchmark datasets (i.e., DES [62], DUT [43], LFSD [75], NJU2K [67], NLPR [64], SIP [70], SSD [76], STERE [77]) for fair comparisons. We split 1485 samples from NJU2K, 800 samples from DUT, and 700 samples from NLPR for training as done in [38, 43, 44, 50, 69, 74]. During training, we use a batch size of 4. All training images are uniformly resized to  $320 \times 320$  and augmented by randomly flipping. The generated  $160 \times 160$  saliency maps are then resized back to the original spatial resolution via bilinear interpolation operation. The parameters of the backbone networks are initialized with the weights of the pretrained ResNet-50, and the rest ones are initialized with a truncated normal. We use the Adam optimizer [78] with a weight decay of  $5e-4$ , and an initial learning rate of  $3e-5$ . The learning rate is divided by 10 after training for 50 epochs. The whole network converges after 70 epochs.

### 4.2 Datasets and evaluation metrics

We conduct extensive experiments on 8 publicly available dataset. The DES [62], also named RGBD135, is a small-scale RGB-D dataset only containing 135 RGB-D images collected by using a Microsoft Kinect. DUT [43] consists of 1200 RGB-D images containing various challenging scenarios (e.g., complex backgrounds, transparent objects and multiple objects). Besides, this dataset is split into a training set of 800 samples and a test set of 400 samples. LFSD [75] includes 100 RGB-D images collected by a Lytro camera. NJU2K [67] is the largest RGB-D dataset that consists of 1985 RGB-D images collected from the Internet and 3D movies. NLPR [64] includes 1000 challenging images collected from indoor and outdoor scenarios, many of which have multiple and small salient objects. SIP [70] is a human-oriented dataset focusing on salient persons in real-world scenarios. It contains 929 high-resolution image pairs captured by Huawei Mate 10. SSD [76] contains 80 images collected from three stereo movies, where the corresponding depth images are obtained by using depth estimation method. STERE [77] consists of 1000 pairs of stereoscopic images downloaded from the Internet.

To provide a comprehensively quantitative evaluation of the performance of the proposed CFIDNet, we adopt eight commonly used standard evaluation metrics, i.e., (1) precision–recall curves (PR-curves), (2) F-measure curves (F-curves), (3) maximum F-measure ( $F_{\beta}^{max}$ ), (4) mean F-



measure ( $F_{\beta}^{avg}$ ), (5) weighted F-measure ( $F_{\beta}^{\omega}$ ), (6) mean absolute error (MAE), (7) S-measure ( $S_{\alpha}$ ), (8) E-measure ( $E_{\xi}$ ).

Basically, the groundtruth is a binary mask, where 0 indicates a background pixel and 1 indicates a salient pixel. However, in a generated saliency map, the saliency confidence of each pixel ranges from 0 to 255. Thus, we first convert saliency maps to binary masks by using a threshold varied from 0 to 255. Then, the average precision and recall scores of images in a dataset can be calculated. After sliding threshold from 0 to 255, a set of precision–recall pairs can be obtained. Based on these pairs of precision–recall scores, we plot the PR-curves.

F-measure is a comprehensive evaluation metric and is defined as a harmonic mean of precision and recall. The F-measure can be calculated as:

$$F_{\beta} = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall}, \quad (22)$$

where  $\beta^2$  is usually set to 0.3 to weigh precision more than recall as suggested in many previous models [27–38]. In this paper, we plot the F-curves and report the maximum F-measure and mean F-measure to provide a more comprehensive evaluation.

The weighted F-measure [79] utilizes weighted precision and weighted recall to construct the evaluation measure, which is defined as:

$$F_{\beta} = \frac{(1 + \beta^2) \times Precision^{\omega} \times Recall^{\omega}}{\beta^2 \times Precision^{\omega} + Recall^{\omega}}, \quad (23)$$

MAE is a simple evaluation metric that is used to reflect the average per-pixel absolute difference between the generated saliency map and the groundtruth. For saliency map  $S$  and groundtruth  $G$ , the MAE between them is computed as:

$$MAE = \frac{1}{H \times W} \sum_{i=1}^H \sum_{j=1}^W |S_{i,j} - G_{i,j}|, \quad (24)$$

where  $H$  and  $W$  are the height and width, respectively.

S-measure [80] is a structure measure that is used to evaluate the structural similarity between a saliency map and the corresponding groundtruth:

$$S_{\alpha} = \alpha \times S_o + (1 - \alpha) \times S_r, \quad (25)$$

where  $\alpha$  is a balance parameter to control the trade-off between  $S_o$  (object-aware structural similarity) and  $S_r$  (region-aware structural similarity). We set it to 0.5 as done in [35, 36].

E-measure [81] is an enhanced-measure that is proposed to compare binary maps. By utilizing both image-level statistics and local pixel-level information, the similarity

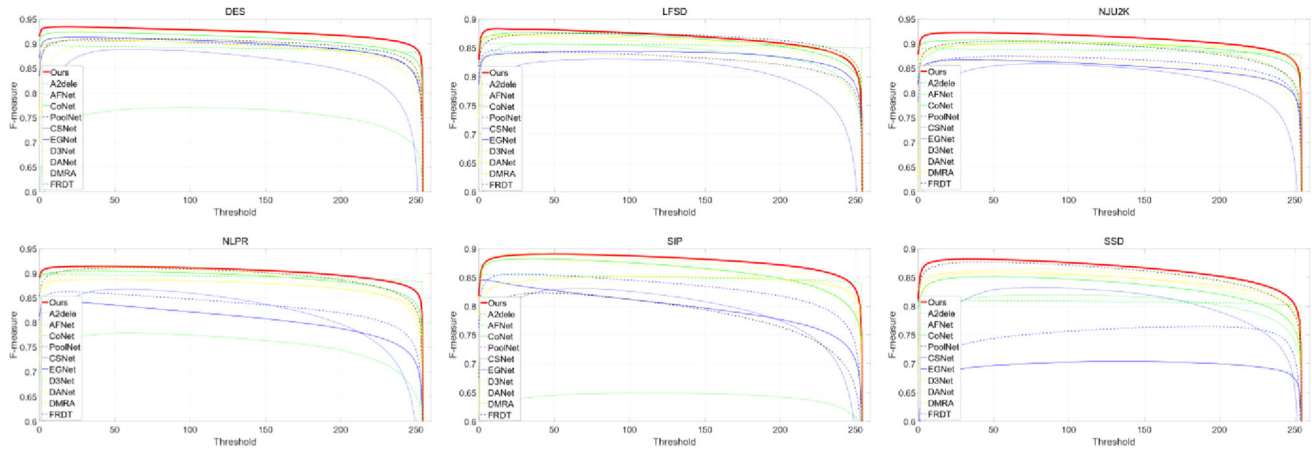
between a saliency map and the groundtruth can be measured.

### 4.3 Comparison with the state-of-the-art

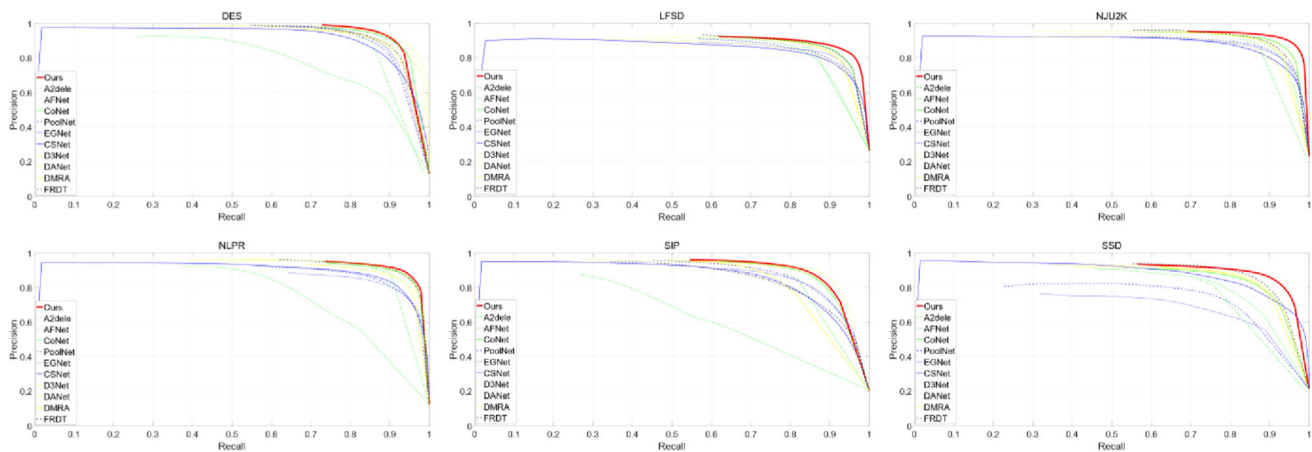
We compare the CFIDNet with 15 state-of-the-art methods including three RGB saliency models (i.e., EGNNet [23], DFI [61] and PoolNet [24]) and twelve RGB-D saliency methods including CTMF [82], PCANet [83], AFNet [39], TANet [84], CFPF [41], DMRA [43], A2dele [38], CoNet [69], DANet [50, 70], FRDT [44], and DCF [85]. For fair comparisons, the saliency maps of the competing methods are provided by their authors or generated by running the publicly available code with pretrained models.

The F-curves and PR-curves of the proposed CFIDNet and 10 state-of-the-art models are shown in Fig. 6 and Fig. 7, respectively. As demonstrated in these two figures, our CFIDNet (shown in red solid line) achieves the best overall performance on almost all evaluation datasets. Besides, to provide a more comprehensive evaluation, the quantitative results of all models in terms of 6 evaluation metrics are shown in Table 1. The table, together with the two figures, demonstrates that the CFIDNet outperforms other state-of-the-art methods on two datasets (i.e., LFSD and SSD) and is also competitive on the remaining datasets.

For a more intuitive comparison of network performance, several representative results of the CFIDNet and other state-of-the-art methods are illustrated in Fig. 8. As is demonstrated in this figure, CFIDNet is able to handle various challenging scenarios. The first row and second row provide the results of images with large salient objects. Compared with other models, the CFIDNet accurately segments the whole salient regions. The third and fourth rows demonstrate the performance of methods in segmenting multiple salient objects. The fifth and sixth rows show the results of images with multi-scale salient objects. It is worth mentioning that in the fifth row, the CFIDNet not only correctly segments the car, but also identifies the left small flag and the right bigger one. Besides, in the sixth row, CFIDNet accurately segments the target that consists of both thin and large structures. These cases validate that the CFIDNet is capable of exploiting multi-scale information for saliency prediction. The seventh and eighth rows are results of images with low contrast scenes, which demonstrate that compared with other methods, the CFIDNet can better leverage the depth information. The ninth and tenth rows show the results of images with complex structures. In summary, the CFIDNet can generate high-quality saliency maps under various complex scenarios.



**Fig. 6** F-measure curves of the proposed CFIDNet and 10 state-of-the-art methods



**Fig. 7** Precision-Recall Curves of the proposed CFIDNet and 10 state-of-the-art methods

#### 4.4 Ablation study

We conduct ablation studies on three aspects to validate the effectiveness of the CFIDNet. Experiments are conducted on four datasets: LFSD, NJU2K, SIP, and SSD.

##### 4.4.1 The number of FIDs

We evaluate the performance of CFIDNet with different number of FIDs. The experimental results are shown in Table 2. We use CFIDNet<sub>*i*</sub> to denote the variant with *i* FIDs. As can be seen from Table 2, the CFIDNet<sub>2</sub> outperforms other variants and achieves the best performance on four benchmark datasets. This is because that by leveraging multiple FIDs, the multi-level features can be iteratively refined to generate finer saliency maps. However, adding excessive number of FIDs may lead to overfitting due to the small number of training samples. Consequently, we select the CFIDNet<sub>2</sub> variant as the final model.

##### 4.4.2 The effectiveness of the FID

To demonstrate the effectiveness of the FID, we conduct ablation experiments. In the baseline network, we replace the FID with a simple decoder, where five  $3 \times 3$  convolutional layers are used to refine multi-level features. Similarly, we stack two simple decoders to process these multi-level features iteratively. Besides, we also implement another two variants to reveal the effectiveness of the non-local block (see *w/o NL* in Table 3) and the tiny U-shape block (see *w/o U* in Table 3) in FRM. As demonstrated in Table 3, the performance degradation of *w/o NL* (i.e.,  $F_{\beta}^{max}$ : 0.7% ~ 4.4%, MAE: 0.001 ~ 0.013,  $S_x$ : 0.4% ~ 3.3%) validates that the non-local block is effective in propagating long-range contextual dependencies and exploiting complementary information between multi-level features. The performance of *w/o U* is also degraded (i.e.,  $F_{\beta}^{max}$ : 1.0% ~ 1.9%, MAE: 0.003 ~ 0.009,  $S_x$ : 0.5% ~ 2.2%), which proves that the tiny U-shape block benefits to excavating multi-scale information.

**Table 1** Comparison of the proposed CFIDNet and 15 state-of-the-art methods on seven datasets in terms of six standard evaluation metrics

Dataset	Metric	CTMF	PCANet	PoolNet	EGNet	AFNet	TANet	CPPP	DFI	DCF	DMRA	A2dele	CoNet	DANet	D <sup>3</sup> Net	FRDT	Ours
LFSD	$F_{\beta}^{max} \uparrow$	0.815	0.829	0.847	0.845	0.856	0.827	0.850	0.858	0.867	0.872	0.858	0.877	0.840	0.849	0.876	<b>0.884</b>
	$F_{\beta}^{avg} \uparrow$	0.781	0.793	0.827	0.830	0.817	0.794	0.813	0.839	0.841	<b>0.849</b>	0.831	0.848	0.805	0.801	0.854	<b>0.857</b>
	$F_{\beta}^o \uparrow$	0.696	0.716	0.769	0.782	0.762	0.719	0.775	0.775	0.783	0.802	0.811	0.799	0.815	0.756	<b>0.814</b>	<b>0.825</b>
	MAE $\downarrow$	0.120	0.112	0.095	0.087	0.091	0.111	0.088	0.088	0.085	0.075	0.076	0.077	0.071	0.095	<b>0.073</b>	<b>0.070</b>
	$S_x \uparrow$	0.796	0.800	0.830	0.838	0.818	0.801	0.828	0.828	0.841	0.841	0.847	0.833	0.862	0.825	<b>0.857</b>	<b>0.870</b>
	$E_{\xi} \uparrow$	0.851	0.856	0.851	0.867	0.868	0.851	0.867	0.867	0.877	0.883	<b>0.899</b>	0.865	<b>0.896</b>	0.853	<b>0.899</b>	0.893
NIU2K	$F_{\beta}^{max} \uparrow$	0.857	0.887	0.874	0.867	0.901	0.888	0.890	0.879	0.919	0.900	0.890	<b>0.907</b>	0.906	0.903	0.905	<b>0.923</b>
	$F_{\beta}^{avg} \uparrow$	0.788	0.844	0.840	0.831	0.871	0.844	0.837	0.843	<b>0.898</b>	<b>0.883</b>	0.868	0.881	0.861	0.840	0.871	0.892
	$F_{\beta}^o \uparrow$	0.720	0.803	0.800	0.797	0.836	0.805	0.828	0.828	0.815	<b>0.882</b>	0.839	<b>0.862</b>	0.850	0.833	0.839	<b>0.882</b>
	MAE $\downarrow$	0.085	0.059	0.067	0.070	0.053	0.061	0.053	0.053	0.062	<b>0.035</b>	0.052	0.050	<b>0.045</b>	0.046	0.049	0.038
	$S_x \uparrow$	0.849	0.877	0.860	0.857	0.884	0.878	0.878	0.878	0.871	0.908	0.881	0.865	<b>0.898</b>	0.897	0.895	<b>0.914</b>
	$E_{\xi} \uparrow$	0.866	0.909	0.880	0.877	<b>0.913</b>	0.909	0.900	0.900	0.877	<b>0.920</b>	<b>0.916</b>	0.899	0.917	0.910	0.909	0.913
NLPR	$F_{\beta}^{max} \uparrow$	0.841	0.864	0.863	0.845	0.778	0.876	0.883	0.880	0.880	0.888	0.898	0.905	0.907	0.904	<b>0.911</b>	0.915
	$F_{\beta}^{avg} \uparrow$	0.724	0.795	0.796	0.769	0.684	0.796	0.818	0.815	0.815	<b>0.890</b>	0.855	0.876	0.869	0.834	0.871	0.885
	$F_{\beta}^o \uparrow$	0.679	0.762	0.779	0.753	0.623	0.780	0.807	0.807	0.804	<b>0.885</b>	0.839	0.857	<b>0.862</b>	0.826	0.857	0.876
	MAE $\downarrow$	0.056	0.044	0.045	0.050	0.070	0.041	0.038	0.038	0.041	<b>0.022</b>	0.031	0.029	0.026	0.030	0.029	0.026
	$S_x \uparrow$	0.860	0.873	0.873	0.863	0.759	0.886	0.884	0.884	0.888	<b>0.924</b>	0.898	0.898	<b>0.915</b>	0.912	0.906	0.914
	$E_{\xi} \uparrow$	0.869	0.916	0.901	0.882	0.852	0.916	0.920	0.920	0.916	<b>0.957</b>	0.942	0.945	0.945	0.944	0.934	<b>0.946</b>
DES	$F_{\beta}^{max} \uparrow$	0.865	0.842	0.907	0.913	0.771	0.853	0.882	0.904	0.910	0.906	0.897	<b>0.924</b>	0.928	0.917	0.910	<b>0.934</b>
	$F_{\beta}^{avg} \uparrow$	0.778	0.774	0.860	0.868	0.724	0.795	0.829	0.855	0.876	0.867	0.865	<b>0.885</b>	<b>0.899</b>	0.876	0.876	0.898
	$F_{\beta}^o \uparrow$	0.686	0.711	0.821	0.831	0.629	0.740	0.787	0.787	0.816	0.857	0.843	0.836	<b>0.867</b>	0.831	0.839	0.875
	MAE $\downarrow$	0.055	0.050	0.035	0.033	0.068	0.046	0.038	0.038	0.039	0.024	0.030	0.028	<b>0.024</b>	0.030	0.030	<b>0.023</b>
	$S_x \uparrow$	0.863	0.843	0.886	0.892	0.768	0.858	0.872	0.872	0.884	0.904	0.899	0.879	<b>0.915</b>	0.904	0.900	0.917
	$E_{\xi} \uparrow$	0.911	0.912	0.922	0.932	0.882	0.919	0.927	0.919	0.919	0.950	0.944	0.917	<b>0.949</b>	0.956	0.940	0.943
SIP	$F_{\beta}^{max} \uparrow$	0.720	0.860	0.856	0.846	0.650	0.851	0.870	0.891	<b>0.899</b>	0.847	0.855	0.883	0.892	0.882	0.823	<b>0.891</b>
	$F_{\beta}^{avg} \uparrow$	0.684	0.825	0.805	0.775	0.588	0.809	0.819	0.836	0.875	0.815	0.827	0.842	<b>0.855</b>	0.831	0.764	0.859
	$F_{\beta}^o \uparrow$	0.535	0.768	0.747	0.728	0.474	0.748	0.788	0.805	0.805	<b>0.839</b>	0.734	0.803	0.822	0.793	0.701	<b>0.816</b>
	MAE $\downarrow$	0.139	0.071	0.079	0.083	0.161	0.075	0.064	0.064	0.056	<b>0.052</b>	0.088	0.070	0.063	0.063	0.088	<b>0.060</b>
	$S_x \uparrow$	0.716	0.842	0.836	0.825	0.628	0.835	0.850	0.850	0.873	<b>0.875</b>	0.800	0.828	0.858	<b>0.875</b>	0.815	<b>0.864</b>
	$E_{\xi} \uparrow$	0.824	0.900	0.883	0.872	0.726	0.894	0.899	0.899	0.912	<b>0.920</b>	0.858	0.889	<b>0.909</b>	0.914	0.903	0.905
SSD	$F_{\beta}^{max} \uparrow$	0.755	0.844	0.764	0.704	0.841	0.834	0.801	0.762	0.867	0.858	0.811	0.851	0.866	<b>0.872</b>	0.877	<b>0.882</b>
	$F_{\beta}^{avg} \uparrow$	0.709	0.786	0.735	0.660	0.813	0.766	0.726	0.720	0.827	0.821	0.777	0.806	0.827	0.793	0.825	<b>0.835</b>
	$F_{\beta}^o \uparrow$	0.622	0.733	0.645	0.576	0.742	0.727	0.708	0.668	0.805	0.787	0.726	0.780	0.795	0.780	<b>0.802</b>	<b>0.819</b>

Table 1 (continued)

Dataset	Metric	CTMF	PCANet	PoolNet	EGNet	AFNet	TANet	CPPP	DFI	DCF	DMRA	A2delete	CoNet	DANet	D <sup>3</sup> Net	FRDT	Ours
	MAE ↓	0.100	0.063	0.110	0.135	0.074	0.063	0.082	0.109	<b>0.050</b>	0.058	0.069	0.059	<b>0.050</b>	0.058	0.053	<b>0.050</b>
	S <sub>z</sub> ↑	0.776	0.842	0.749	0.707	0.820	0.839	0.807	0.770	0.864	0.856	0.797	0.853	0.864	<b>0.866</b>	0.872	<b>0.879</b>
	E <sub>z</sub> ↑	0.838	0.890	0.845	0.788	0.886	0.886	0.832	0.824	0.898	0.898	0.850	0.896	<b>0.908</b>	0.892	<b>0.904</b>	<b>0.908</b>
STERE	F <sub>β</sub> <sup>max</sup> ↑	0.848	0.875	0.872	0.867	0.848	0.878	0.889	0.887	<b>0.911</b>	0.802	0.892	0.909	0.906	0.897	0.839	<b>0.908</b>
	F <sub>β</sub> <sup>avg</sup> ↑	0.771	0.826	0.841	0.829	0.807	0.835	0.830	0.852	<b>0.884</b>	0.762	0.875	<b>0.884</b>	0.868	0.833	0.783	0.879
	F <sub>β</sub> <sup>o</sup> ↑	0.698	0.778	0.787	0.785	0.752	0.787	0.817	0.815	<b>0.868</b>	0.647	0.847	0.866	0.846	0.815	0.729	<b>0.861</b>
	MAE ↓	0.086	0.064	0.063	0.065	0.075	0.060	0.051	0.054	0.039	0.087	<b>0.043</b>	<b>0.037</b>	<b>0.043</b>	0.054	0.077	<b>0.043</b>
	S <sub>z</sub> ↑	0.848	0.875	0.856	0.853	0.825	0.871	0.879	0.877	0.902	0.752	0.876	<b>0.905</b>	<b>0.901</b>	0.891	0.841	<b>0.901</b>
	E <sub>z</sub> ↑	0.870	0.907	0.899	0.894	0.887	0.916	0.907	0.911	<b>0.929</b>	0.816	0.913	0.927	0.921	0.911	0.871	<b>0.924</b>

The top three results are highlighted in bold, italic, and bolditalic, respectively

#### 4.4.3 Cross-modality feature integration strategy

To verify the effectiveness of our cross-modality feature integration strategy, we compare it with other three methods. The integration process of our methods can be formulated as:

$$f_i = f_r + C_3(\text{cat}(f_r, f_d)), \quad (26)$$

where  $f_i$  is the integrated feature. We implement three models with different integration strategies. The first one is named *Add.*, the integration process of which is defined as:

$$f_i = f_r + C_3(f_r + f_d). \quad (27)$$

The second one is named *Mul.*, the integration process of which can be formulated as:

$$f_i = f_r + C_3(\text{sigmoid}(f_r) \times f_d). \quad (28)$$

As demonstrated in Table 5, our integration strategy shows better performance on these evaluation datasets, which validates the effectiveness of the proposed integration strategy.

The third one is named *Cat.*, where the attention block is applied on the concatenated RGB and depth feature maps. The quantitative experimental results are shown in Table 4. As demonstrated in the table, CFIDNet outperforms *Cat.* by a non-negligible margin. The underlying reason is that in many cases, RGB features are more informative than depth ones. Thus, the channel attention operation applied on the concatenated features may only highlight channels belonging to RGB features and makes the whole model bias its learning toward only RGB knowledge, while informative depth cues may be overlooked.

#### 4.4.4 Computational complexity

To fully compare the CFIDNet with other state-of-the-art methods, we report the computational complexities of CFIDNet and 9 existing high-performance approaches including A2delete, CoNet, CSNet, D<sup>3</sup>Net, DANet, DMRA, EGNet, FRDT, PoolNet. For fair comparisons, each model takes 320 × 320 images as inputs. Experiments are conducted 20 times on a machine with a NVIDIA Titan Xp GPU. The average inference speed of these methods is shown in Table 5.

## 5 Conclusion

In this paper, we design a novel deep neural network named CFIDNet for RGB-D salient object detection. First, we propose the FEM to excavate informative cues from depth modality and exploit complementary information between depth and RGB features. Thus, RGB feature can

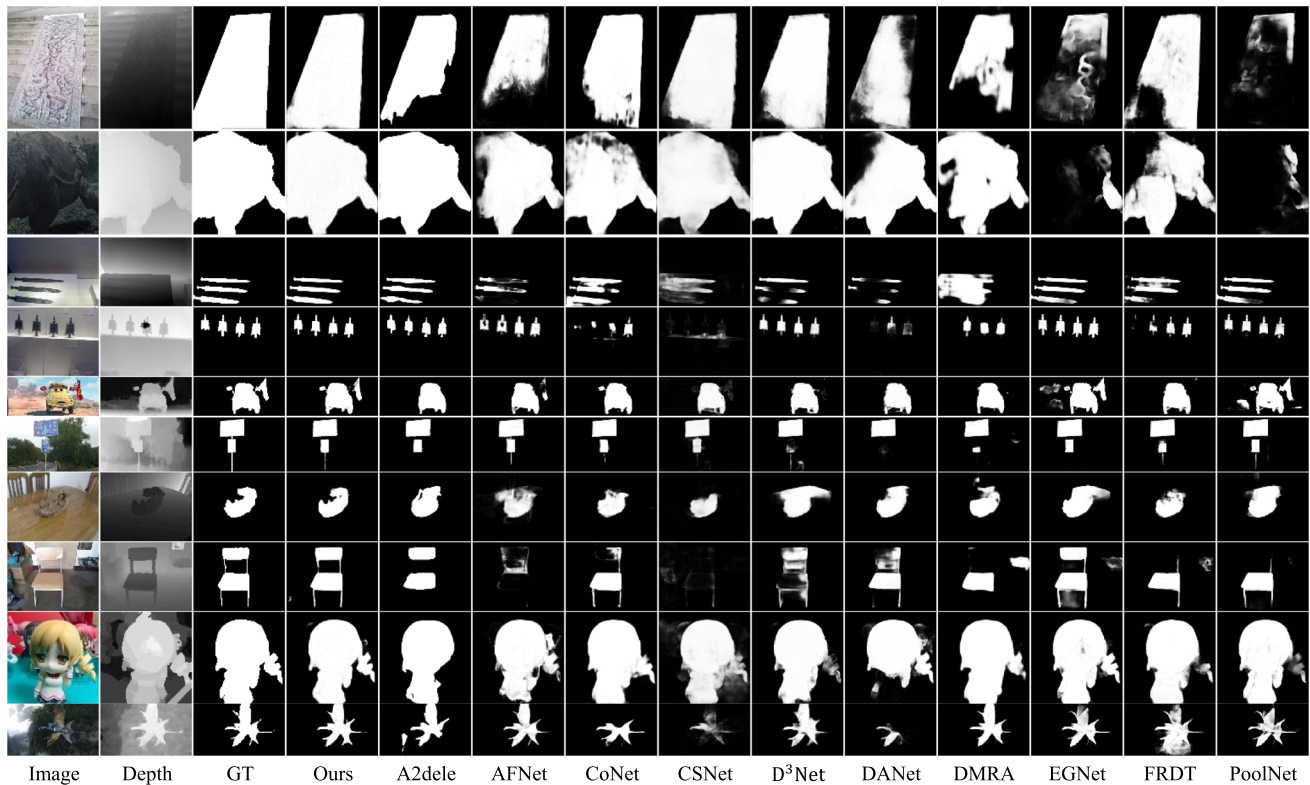


Fig. 8 Qualitative comparisons of the proposed CFIDNet and 10 state-of-the-art methods

Table 2 Comparisons of the CFIDNet with different number of FIDs on LFSD, NJU2K, SIP, and SSD datasets

	Complexity			LFSD			NJU2K			SIP			SSD		
	Param	FLOPs	Time(s)	$F_{\beta}^{max}$ ↑	MAE ↓	$S_z$ ↑	$F_{\beta}^{max}$ ↑	MAE ↓	$S_z$ ↑	$F_{\beta}^{max}$ ↑	MAE ↓	$S_z$ ↑	$F_{\beta}^{max}$ ↑	MAE ↓	$S_z$ ↑
CFIDNet <sub>1</sub>	51.07 M	36.82G	0.037	0.856	0.082	0.847	0.917	0.040	0.909	0.870	0.068	0.850	0.862	0.057	0.855
CFIDNet <sub>4</sub>	59.46 M	55.11G	0.069	0.875	0.073	0.862	0.913	0.042	0.907	0.872	0.071	0.843	0.854	0.061	0.859
CFIDNet <sub>2</sub>	53.86 M	42.92G	0.046	<b>0.884</b>	<b>0.070</b>	<b>0.870</b>	<b>0.923</b>	<b>0.038</b>	<b>0.914</b>	<b>0.891</b>	<b>0.060</b>	<b>0.864</b>	<b>0.882</b>	<b>0.050</b>	<b>0.879</b>

The best results are highlighted in Boldface. The FLOPs and inference speed of different variants are calculated with  $320 \times 320$  images

Table 3 Ablation analyses for the network architecture on LFSD, NJU2K, SIP, and SSD datasets in terms of three evaluation metrics

	Complexity			LFSD			NJU2K			SIP			SSD		
	Param	FLOPs	Time(s)	$F_{\beta}^{max}$ ↑	MAE ↓	$S_z$ ↑	$F_{\beta}^{max}$ ↑	MAE ↓	$S_z$ ↑	$F_{\beta}^{max}$ ↑	MAE ↓	$S_z$ ↑	$F_{\beta}^{max}$ ↑	MAE ↓	$S_z$ ↑
baseline	48.64 M	33.18G	0.025	0.629	0.211	0.667	0.771	0.126	0.786	0.352	0.312	0.478	0.518	0.241	0.568
w/o U	52.53 M	42.66G	0.040	0.872	0.073	0.865	0.913	0.041	0.908	0.873	0.069	0.856	0.863	0.059	0.857
w/o NL	51.59 M	36.57G	0.044	0.840	0.084	0.837	0.916	0.039	0.910	0.878	0.064	0.859	0.848	0.061	0.850
CFIDNet	53.86 M	42.92G	0.046	<b>0.884</b>	<b>0.070</b>	<b>0.870</b>	<b>0.923</b>	<b>0.038</b>	<b>0.914</b>	<b>0.891</b>	<b>0.060</b>	<b>0.864</b>	<b>0.882</b>	<b>0.050</b>	<b>0.879</b>

The best results are highlighted in Boldface. Baseline: replacing each FID with 5 convolutional layers; w/o U: removing the tiny U-shape block in FRM; w/o NL: removing the non-local block in FRM. The FLOPs and inference speed of different variants are calculated with  $320 \times 320$  images

**Table 4** Ablation analyses for the multi-modality feature integration strategy

	Complexity			LFSD			NJU2K			SIP			SSD		
	Param	FLOPs	Time(s)	$F_{\beta}^{max} \uparrow$	MAE $\downarrow$	$S_z \uparrow$	$F_{\beta}^{max} \uparrow$	MAE $\downarrow$	$S_z \uparrow$	$F_{\beta}^{max} \uparrow$	MAE $\downarrow$	$S_z \uparrow$	$F_{\beta}^{max} \uparrow$	MAE $\downarrow$	$S_z \uparrow$
Add	53.68 M	41.65G	0.046	0.868	0.075	0.861	0.916	0.039	0.911	0.883	0.062	0.863	0.852	0.056	0.854
Mul	53.67 M	41.65G	0.046	0.862	0.080	0.858	0.902	0.047	0.900	0.865	0.073	0.850	0.821	0.082	0.822
Cat	53.89 M	42.92G	0.046	0.867	0.073	0.863	0.918	0.039	0.908	0.886	0.063	0.861	0.861	0.053	0.864
CFIDNet	53.86 M	42.92G	0.046	<b>0.884</b>	<b>0.070</b>	<b>0.870</b>	<b>0.923</b>	<b>0.038</b>	<b>0.914</b>	<b>0.891</b>	<b>0.060</b>	<b>0.864</b>	<b>0.882</b>	<b>0.050</b>	<b>0.879</b>

Add.: integration by element-wise addition; Mul.: integration by element-wise multiplication. Cat.: attention block is applied on the concatenated RGB and depth feature maps. The FLOPs and inference speed of different variants are calculated with  $320 \times 320$  images

**Table 5** Average inference speed comparisons between CFIDNet and 9 state-of-the-art models

Method	A2dele	CoNet	CSNet	D <sup>3</sup> Net	DANet
Time(s)	0.0079	0.0251	0.0279	0.0280	0.0168
Method	DMRA	EGNet	FRDT	PoolNet	Ours
Time(s)	0.0456	0.0920	0.0665	0.0249	0.0462

be enhanced after combined with the exploited complementary information. Besides, we take into account the level-specific characteristics of features extracted from the backbone, and propose the FRM. The FRM is effective in capturing global contextual dependencies and exploiting multi-scale information. By leveraging multiple FRMs sequentially, our FID is able to refine multi-level features. Afterward, CFID is proposed to refine features of all levels iteratively. Hence, the CFIDNet can accurately segment salient objects. Experimental results on 7 widely used benchmark datasets validate that the CFIDNet is competitive compared with 15 state-of-the-art counterparts in terms of eight evaluation metrics.

**Acknowledgements** This work was supported by the National Natural Science Foundation of China (under Grant 51807003).

## Declarations

**Conflict of interest** We declare that we have no financial and personal relationships with other people or organizations that can inappropriately influence our work; there is no professional or other personal interest of any nature or kind in any product, service and/or company that could be construed as influencing the position presented in, or the review of, the manuscript entitled.

## References

- Borji A, Cheng MM, Jiang H, Li J (2015) Salient object detection: a benchmark. *IEEE Trans Image Process.* <https://doi.org/10.1109/TIP.2015.2487833>
- Wang W, Lai Q, Fu H, Shen J, Ling H, Yang R (2021) Salient object detection in the deep learning era: an in-depth survey. *IEEE Trans Pattern Anal Mach Intell.* <https://doi.org/10.1109/TPAMI.2021.3051099>
- Cheng MM, Liu Y, Lin WY, Zhang Z, Rosin PL, Torr PHS (2019) BING: Binarized normed gradients for objectness estimation at 300fps. *Comput Vis Media.* <https://doi.org/10.1007/s41095-018-0120-1>
- Cheng MM, Zhang FL, Mitra NJ, Huang X, Hu SM (2010) RepFinder: Finding approximately repeated scene elements for image editing. *ACM Trans Graph.* <https://doi.org/10.1145/1778765.1778820>
- Liu C et al (2020) Aggregation signature for small object tracking. *IEEE Trans Image Process.* <https://doi.org/10.1109/TIP.2019.2940477>
- Borji A, Frintrop S, Sihite DN, Itti L (2012) Adaptive object tracking by learning background context. In: 2012 IEEE computer society conference on computer vision and pattern recognition workshops. pp 23–30. *IEEE.* <https://doi.org/10.1109/CVPRW.2012.6239191>
- Hong S, You T, Kwak S, Han B (2015) Online tracking by learning discriminative saliency map with convolutional neural network. In: 32nd international conference on machine learning, ICML 2015, vol 1
- Zhao R, Ouyang W, Wang X (2013) Unsupervised salience learning for person re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition <https://doi.org/10.1109/CVPR.2013.460>
- Fan DP, Wang W, Cheng MM, Shen J (2019) Shifting more attention to video salient object detection. In: Proceedings of the IEEE computer society conference on computer vision and pattern recognition, vol 2019-June, <https://doi.org/10.1109/CVPR.2019.00875>
- Yan P et al (2019) Semi-supervised video salient object detection using pseudo-labels. In: Proceedings of the IEEE international conference on computer vision, vol 2019-October, <https://doi.org/10.1109/ICCV.2019.00738>
- Wang W, Shen J, Yu Y, Ma KL (2017) Stereoscopic thumbnail creation via efficient stereo saliency detection. *IEEE Trans Vis Comput Graph.* <https://doi.org/10.1109/TVCG.2016.2600594>
- Cheng MM, Mitra NJ, Huang X, Torr PHS, Hu SM (2015) Global contrast based salient region detection. *IEEE Trans Pattern Anal Mach Intell.* <https://doi.org/10.1109/TPAMI.2014.2345401>

13. Xiao F, Peng L, Fu L, Gao X (2018) Salient object detection based on eye tracking data. *Signal Process.* <https://doi.org/10.1016/j.sigpro.2017.10.019>
14. Jiang H, Wang J, Yuan Z, Wu Y, Zheng N, Li S (2013) Salient object detection: A discriminative regional feature integration approach. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp 2083–2090. <https://doi.org/10.1109/CVPR.2013.271>
15. Zhang J, Ehinger KA, Wei H, Zhang K, Yang J (2017) A novel graph-based optimization framework for salient object detection. *Pattern Recognit.* <https://doi.org/10.1016/j.patcog.2016.10.025>
16. Lu S, Lim JH (2012) Saliency modeling from image histograms. In: *European Conference on Computer Vision*, pp 312–332. Springer, Berlin 2012.
17. Klein DA, Frintrop S (2011) Center-surround divergence of feature statistics for salient object detection. In: *2011 international conference on computer vision*. IEEE, <https://doi.org/10.1109/ICCV.2011.6126499>
18. Chen T, Hu X, Xiao J, Zhang G (2021) BPFINet: boundary-aware progressive feature integration network for salient object detection. *Neurocomputing.* <https://doi.org/10.1016/j.neucom.2021.04.078>
19. Tu Z, Ma Y, Li C, Li C, Tang J, Luo B (2020) Edge-guided non-local fully convolutional network for salient object detection. *IEEE Trans Circuits Syst Video Technol.* <https://doi.org/10.1109/tcsvt.2020.2980853>
20. Hou Q, Cheng MM, Hu X, Borji A, Tu Z, Torr PHS (2019) Deeply supervised salient object detection with short connections. *IEEE Trans Pattern Anal Mach Intell.* <https://doi.org/10.1109/TPAMI.2018.2815688>
21. Luo Z, Mishra A, Achkar A, Eichel J, Li S, Jodoin PM (2017) Non-local deep features for salient object detection. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*. pp 6609–6617. <https://doi.org/10.1109/CVPR.2017.698>
22. Zhang P, Wang D, Lu H, Wang H, Yin B (2017) Learning uncertain convolutional features for accurate saliency detection. In: *Proceedings of the IEEE international conference on computer vision*, vol 2017-October, <https://doi.org/10.1109/ICCV.2017.32>
23. Zhao J, Liu JJ, Fan DP, Cao Y, Yang J, Cheng MM (2019) EGNet: Edge guidance network for salient object detection. In: *Proceedings of the IEEE international conference on computer vision*, vol 2019-October, <https://doi.org/10.1109/ICCV.2019.00887>
24. Liu JJ, Hou Q, Cheng MM, Feng J, Jiang J (2019) A simple pooling-based design for real-time salient object detection, <https://doi.org/10.1109/CVPR.2019.00404>
25. Zhang L, Dai J, Lu H, He Y, Wang G (2018) A bi-directional message passing model for salient object detection, <https://doi.org/10.1109/CVPR.2018.00187>
26. Zhu L et al (2020) Aggregating attentional dilated features for salient object detection. *IEEE Trans Circuits Syst Video Technol.* <https://doi.org/10.1109/TCSVT.2019.2941017>
27. Qin X, Zhang Z, Huang C, Gao C, Dehghan M, Jagersand M (2019) Basnet: boundary-aware salient object detection. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, vol 2019-June, <https://doi.org/10.1109/CVPR.2019.00766>
28. Wu Z, Su L, Huang Q (2019) Cascaded partial decoder for fast and accurate salient object detection. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, vol 2019-June, <https://doi.org/10.1109/CVPR.2019.00403>
29. Wei J, Wang S, Huang Q (2020) F3Net: fusion, feedback and focus for salient object detection. In: *Proceedings of the AAAI Conference on Artificial Intelligence*, vol 34, no 07 SE-AAAI Technical Track: Vision, pp 12321–12328, <https://doi.org/10.1609/aaai.v34i07.6916>
30. Gao S-H, Tan Y-Q, Cheng M-M, Lu C, Chen Y, Yan S (2020) Highly efficient salient object detection with 100K parameters. In: *Computer Vision—ECCV 2020*, pp 702–721
31. Zhou H, Xie X, Lai J-H, Chen Z, Yang L (2020) Interactive two-stream decoder for accurate and fast saliency detection, doi: <https://doi.org/10.1109/cvpr42600.2020.00916>
32. Pang Y, Zhao X, Zhang L, Lu H (2020) Multi-scale interactive network for salient object detection. *CVPR.* <https://doi.org/10.1109/cvpr42600.2020.00943>
33. Deng Z et al (2018) R3Net: recurrent residual refinement network for saliency detection. In: *IJCAI international joint conference on artificial intelligence*, vol 2018-July, <https://doi.org/10.24963/ijcai.2018/95>
34. Qin X, Zhang Z, Huang C, Dehghan M, Zaiane OR, Jagersand M (2020) U2-Net: Going deeper with nested U-structure for salient object detection. *Pattern Recognit.* <https://doi.org/10.1016/j.patcog.2020.107404>
35. Zhai Y et al (2020) Bifurcated backbone strategy for RGB-D salient object detection, arXiv. 2020
36. Chen Z, Cong R, Xu Q, Huang Q (2020) DPANet: depth potentiality-aware gated attention network for RGB-D salient object detection. *IEEE Trans Image Process.* <https://doi.org/10.1109/tip.2020.3028289>
37. Da Jin W, Xu J, Han Q, Zhang Y, Cheng MM (2021) CDNet: complementary depth network for rgb-d salient object detection. *IEEE Trans Image Process.* <https://doi.org/10.1109/TIP.2021.3060167>
38. Piao Y, Rong Z, Zhang M, Ren W, Lu H (2020) A2dele: adaptive and attentive depth distiller for efficient RGB-D salient object detection, <https://doi.org/10.1109/CVPR42600.2020.00908>
39. Wang N, Gong X (2019) Adaptive fusion for RGB-D salient object detection. *IEEE Access* 7:55277–55284. <https://doi.org/10.1109/ACCESS.2019.2913107>
40. Zhang M, Fei SX, Liu J, Xu S, Piao Y, Lu H (2020) Asymmetric two-stream architecture for accurate RGB-D saliency detection. *ECCV.* [https://doi.org/10.1007/978-3-030-58604-1\\_23](https://doi.org/10.1007/978-3-030-58604-1_23)
41. Zhao JX, Cao Y, Fan DP, Cheng MM, Li XY, Zhang L (2019) Contrast prior and fluid pyramid integration for rgb-d salient object detection. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, vol 2019-June, <https://doi.org/10.1109/CVPR.2019.00405>
42. Li G, Liu Z, Ye L, Wang Y, Ling H (2020) Cross-modal weighting network for RGB-D salient object detection. In: *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, 2020, vol 12362 LNCS, [https://doi.org/10.1007/978-3-030-58520-4\\_39](https://doi.org/10.1007/978-3-030-58520-4_39)
43. Piao Y, Ji W, Li J, Zhang M, Lu H (2019) Depth-induced multi-scale recurrent attention network for saliency detection. In: *Proceedings of the IEEE international conference on computer vision*, vol 2019-October, <https://doi.org/10.1109/ICCV.2019.00735>
44. Zhang M, Zhang Y, Piao Y, Hu B, Lu H (2020) Feature reintegration over differential treatment: a top-down and adaptive fusion network for RGB-D salient object detection, <https://doi.org/10.1145/3394171.3413969>
45. Pang Y, Zhang L, Zhao X, Lu H (2020) Hierarchical dynamic filtering network for rgb-d salient object detection. *ECCV.* [https://doi.org/10.1007/978-3-030-58595-2\\_15](https://doi.org/10.1007/978-3-030-58595-2_15)
46. Li G, Liu Z, Ling H (2020) ICNet: information conversion network for RGB-D based salient object detection". *IEEE Trans Image Process.* <https://doi.org/10.1109/TIP.2020.2976689>

47. Wu J, Zhou W, Luo T, Yu L, Lei J (2021) Multiscale multilevel context and multimodal fusion for RGB-D salient object detection. *Signal Process.* <https://doi.org/10.1016/j.sigpro.2020.107766>
48. Fu K, Fan DP, Ji GP, Zhao Q (2020) JL-DCF: joint learning and densely-cooperative fusion framework for RGB-D salient object detection. <https://doi.org/10.1109/CVPR42600.2020.00312>
49. Chen H, Li Y, Su D (2019) Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. *Pattern Recognit.* <https://doi.org/10.1016/j.patcog.2018.08.007>
50. Zhao X, Zhang L, Pang Y, Lu H, Zhang L (2020) A single stream network for robust and real-time RGB-D salient object detection. *ECCV.* [https://doi.org/10.1007/978-3-030-58542-6\\_39](https://doi.org/10.1007/978-3-030-58542-6_39)
51. Li G, Liu Z, Chen M, Bai Z, Lin W, Ling H (2021) Hierarchical alternate interaction network for rgb-d salient object detection. *IEEE Trans Image Process.* <https://doi.org/10.1109/TIP.2021.3062689>
52. Wang X, Girshick R, Gupta A, He K (2018) Non-local Neural Networks. <https://doi.org/10.1109/CVPR.2018.00813>
53. Lu S, Tan C, Lim JH (2014) Robust and efficient saliency modeling from image co-occurrence histograms. *IEEE Trans Pattern Anal Mach Intell.* <https://doi.org/10.1109/TPAMI.2013.158>
54. Zhang P, Wang D, Lu H, Wang H, Ruan X (2017) Amulet: aggregating multi-level convolutional features for salient object detection. In: *Proceedings of the IEEE international conference on computer vision*, vol 2017-October, <https://doi.org/10.1109/ICCV.2017.31>
55. Wang T, Borji A, Zhang L, Zhang P, Lu H (2017) A stagewise refinement model for detecting salient objects in images. In: *Proceedings of the IEEE international conference on computer vision*, vol 2017-October, <https://doi.org/10.1109/ICCV.2017.433>
56. Zhao H, Shi J, Qi X, Wang X, Jia J (2017) Pyramid scene parsing network. In: *Proceedings - 30th IEEE conference on computer vision and pattern recognition*, CVPR 2017, vol 2017-January, <https://doi.org/10.1109/CVPR.2017.660>
57. Liu N, Han J, Yang MH (2018) PiCANet: learning pixel-wise contextual attention for saliency detection. <https://doi.org/10.1109/CVPR.2018.00326>
58. Feng M, Lu H, Ding E (2019) Attentive feedback network for boundary-aware salient object detection. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, vol 2019-June, <https://doi.org/10.1109/CVPR.2019.00172>
59. Zhao T, Wu X (2019) Pyramid feature attention network for saliency detection. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, vol 2019-June, <https://doi.org/10.1109/CVPR.2019.00320>
60. Wu Z, Su L, Huang Q (2019) Stacked cross refinement network for edge-aware salient object detection. <https://doi.org/10.1109/ICCV.2019.00736>
61. Liu JJ, Hou Q, Cheng MM (2020) Dynamic feature integration for simultaneous detection of salient object, edge, and skeleton. *IEEE Trans Image Process.* <https://doi.org/10.1109/TIP.2020.3017352>
62. Cheng Y, Fu H, Wei X, Xiao J, Cao X (2014) Depth enhanced saliency detection method. <https://doi.org/10.1145/2632856.2632866>
63. Zhu C, Li G, Wang W, Wang R (2017) An innovative salient object detection using center-dark channel prior. In: *Proceedings - 2017 IEEE international conference on computer vision workshops, ICCVW 2017*, vol 2018-January, <https://doi.org/10.1109/ICCVW.2017.178>
64. Peng H, Li B, Xiong W, Hu W, Ji R (2014) RGBD salient object detection: A benchmark and algorithms. *ECCV.* [https://doi.org/10.1007/978-3-319-10578-9\\_7](https://doi.org/10.1007/978-3-319-10578-9_7)
65. Song H, Liu Z, Du H, Sun G, Le Meur O, Ren T (2017) Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. *IEEE Trans Image Process.* <https://doi.org/10.1109/TIP.2017.2711277>
66. Feng D, Barnes N, You S, McCarthy C (2016) Local background enclosure for RGB-D salient object detection. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, vol 2016-December, <https://doi.org/10.1109/CVPR.2016.257>
67. Ju R, Ge L, Geng W, Ren T, Wu G (2014) Depth saliency based on anisotropic center-surround difference. <https://doi.org/10.1109/ICIP.2014.7025222>
68. Zhu C, Cai X, Huang K, Li TH, Li G (2019) PDNet: prior-model guided depth-enhanced network for salient object detection. In: *Proceedings - IEEE international conference on multimedia and expo*, vol 2019-July, <https://doi.org/10.1109/ICME.2019.00042>
69. Ji W, Li J, Zhang M, Piao Y, Lu H (2020) Accurate RGB-D salient object detection via collaborative learning. In: *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)*, vol 12363 LNCS, [https://doi.org/10.1007/978-3-030-58523-5\\_4](https://doi.org/10.1007/978-3-030-58523-5_4)
70. Fan D-P, Lin Z, Zhang Z, Zhu M, Cheng M-M (2020) Rethinking RGB-D salient object detection: models, data sets, and large-scale benchmarks. *IEEE Trans Neural Netw Learn Syst.* <https://doi.org/10.1109/TNNLS.2020.2996406>
71. He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: *Proceedings of the IEEE computer society conference on computer vision and pattern recognition*, vol 2016-December, <https://doi.org/10.1109/CVPR.2016.90>
72. Gupta S, Girshick R, Arbeláez P, Malik J (2014) Learning rich features from RGB-D images for object detection and segmentation. *ECCV.* [https://doi.org/10.1007/978-3-319-10584-0\\_23](https://doi.org/10.1007/978-3-319-10584-0_23)
73. Krähenbühl P, Koltun V (2012) Efficient inference in fully connected CRFs with gaussian edge potentials. *Adv Neural Inf Process Syst* 24:109–117
74. Liu N, Zhang N, Han J (2020) Learning selective self-mutual attention for RGB-D saliency detection. <https://doi.org/10.1109/CVPR42600.2020.01377>
75. Li N, Ye J, Ji Y, Ling H, Yu J (2017) Saliency detection on light field. *IEEE Trans Pattern Anal Mach Intell.* <https://doi.org/10.1109/TPAMI.2016.2610425>
76. Li G, Zhu C (2017) A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In: *Proceedings - 2017 IEEE international conference on computer vision workshops, ICCVW 2017*, vol 2018-January, <https://doi.org/10.1109/ICCVW.2017.355>
77. Niu Y, Geng Y, Li X, Liu F (2012) Leveraging stereopsis for saliency analysis. <https://doi.org/10.1109/CVPR.2012.6247708>
78. Kingma DP, Ba JL (2015) Adam: a method for stochastic optimization
79. Margolin R, Zelnik-Manor L, Tal A (2014) How to evaluate foreground maps. <https://doi.org/10.1109/CVPR.2014.39>
80. Fan DP, Cheng MM, Liu Y, Li T, Borji A (2017) Structure-measure: a new way to evaluate foreground maps. In: *Proceedings of the IEEE international conference on computer vision*, vol 2017-October, <https://doi.org/10.1109/ICCV.2017.487>
81. Fan DP, Gong C, Cao Y, Ren B, Cheng MM, Borji A (2018) Enhanced-alignment measure for binary foreground map evaluation. In: *IJCAI International joint conference on artificial intelligence*, vol 2018-July, <https://doi.org/10.24963/ijcai.2018/97>
82. Han J, Chen H, Liu N, Yan C (2018) Li X “CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion.” *IEEE Trans Cybern.* <https://doi.org/10.1109/TCYB.2017.2761775>



83. Chen H, Li Y (2018) Progressively complementarity-aware fusion network for RGB-D salient object detection, <https://doi.org/10.1109/CVPR.2018.00322>
84. Chen H, Li Y (2019) Three-stream attention-aware network for rgb-d salient object detection. IEEE Trans Image Process. <https://doi.org/10.1109/TIP.2019.2891104>
85. Ji W et al (2021) Calibrated RGB-D salient object detection. In: CVPR, pp 9471–9481
86. Hussain T, Anwar S, Ullah A, Muhammad K, Baik SW (2021) Densely deformable efficient salient object detection network, In: CoRR, vol abs/2102.06407, [Online]. Available: <https://arxiv.org/abs/2102.06407>
87. Pang J, Chen K, Shi J, Feng H, Ouyang W, Lin D (2019) Libra R-CNN: towards balanced learning for object detection. CVPR. <https://doi.org/10.1109/CVPR.2019.00091>

**Publisher's Note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.