Pattern Recognition 128 (2022) 108666

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Encoder deep interleaved network with multi-scale aggregation for RGB-D salient object detection

Guang Feng, Jinyu Meng, Lihe Zhang*, Huchuan Lu

School of Information and Communication Engineering, Dalian University of Technology, Dalian 116023, China

ARTICLE INFO

Article history:
Received 7 January 2021
Revised 10 March 2022
Accepted 22 March 2022
Available online 24 March 2022

Keywords:
RGB-D salient object detection
Deep interleaved encoder
Cross-modal mutual guidance
Residual multi-scale feature aggregation
Real-time

ABSTRACT

Recently, RGB-D salient object detection (SOD) has aroused widespread research interest. Existing RGB-D SOD approaches mainly consider the cross-modal information fusion in the decoder. And their multi-modal interaction mainly concentrates on the same level of features between RGB stream and depth stream. They do not deeply explore the coherence of multi-model features at different levels. In this paper, we design a two-stream deep interleaved encoder network to extract RGB and depth information and realize their mixing simultaneously. This network allows us to gradually learn multi-modal representation at different levels from shallow to deep. Moreover, to further fuse multi-modal features in the decoding stage, we propose a cross-modal mutual guidance module and a residual multi-scale aggregation module to implement the global guidance and local refinement of the salient region. Extensive experiments on six benchmark datasets demonstrate that the proposed approach performs favorably against most state-of-the-art methods under different evaluation metrics. During the testing stage, this model can run at a real-time speed of 93 FPS.

1. Introduction

Salient object detection (SOD) aims to identify the most distinct regions or objects in an image by imitating the characteristics of human visual attention. SOD allocates limited computing resources to important regions in the scene, which not only provides effective information for the subsequent visual tasks but also can eliminate the interference of redundant regions and reduce the computational cost. Due to this advantages of the SOD task, it is beneficial for various computer vision tasks, including image captioning [1], image retrieval [2], semantic segmentation [3], object proposal [4], and visual tracking [5,6] etc.

With the development of fully convolutional networks (FCNs), many FCN-based SOD methods have emerged [7–9], and they have pushed the performance of SOD to a new level. Most of the above methods use RGB images as the input of the network, but when they are faced with low contrast or cluttered scenes, the detection results are still unsatisfactory. Recently, some saliency detection methods [10–16] introduce depth information into the network, and they provide an intuitive spatial structure representation for the encoder as a supplement (as shown in Fig. 1). Therefore, the RGB-D based SOD methods gradually broke through the performance bottleneck of the RGB-based methods. Existing methods can be roughly divided into three categories: input fusion [17,18], result fusion [19,20] and feature fusion [21–23]. Most of methods focus on the feature fusion strategy, that is, to interact with the intermediate feature information of the two modalities. Specifically, they either utilize the fusion results of depth and RGB features to generate the final prediction [22,24], or exploit the depth information as a guide to update the RGB features [13,25]. These feature fusion-based RGB-D salient object detection methods usually employ two independent feature encoders to extract RGB and depth features, and then use the feature pyramid network (FPN) to fuse two modality features from deep layer to shallow layer in the decoder. Thus, the highest-level cross-modal features generated by this paradigm directly come from two deepest single-modality encoder features, rather than from shallow multi-modal fused features. This strategy ignores the continuity and coherence of multi-modal information propagation from shallow layer to deep layer.

In this work, we design a two-stream deep interleaved backbone (DIB) to realize the bottom-up cross-modality interaction in the encoder. The bottom-up fusion can more easily align two modalities. In the decoder, we employ a residual multi-scale aggregation module (RMSA) to perform the top-down multi-modal interaction from deep layer to shallow layer. As a result, the multi-modal feature fusion process runs through the entire network, and the bottom-up and top-down feature interactions jointly promote saliency prediction. Besides, we also present a cross-modal mutual

* Corresponding author.
E-mail address: zhanglihe@dlut.edu.cn (L. Zhang).

https://doi.org/10.1016/j.patcog.2022.108666
0031-3203/

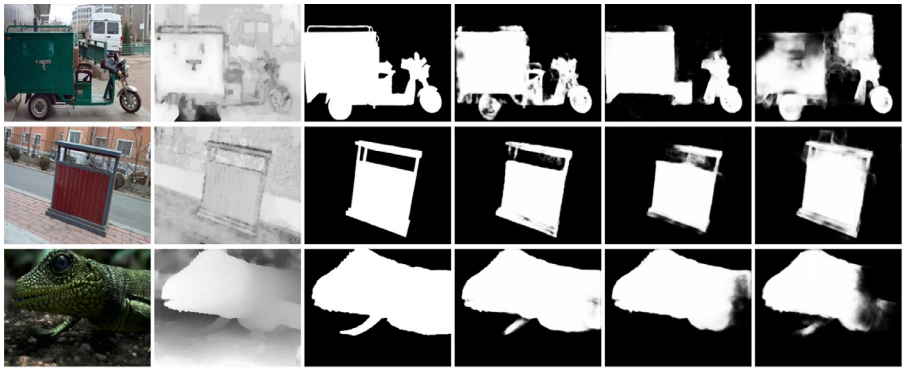


Fig. 1. Visual comparison with other methods. From left to right: RGB image, depth image, ground truth, saliency maps of ours, DANet [18], and D3Net [26].

guidance module to model the interdependencies between channels to assist the localization of the salient region.

Our main contributions are as follows:

- We propose a novel deep interleaved backbone (DIB) to gradually realize the effective transmission of multi-modal information between each convolution stage. This backbone bridges the gap between the two streams and produces a compact multi-modal representation.
- We propose a cross-modal mutual guidance module (CMG) and a residual multi-scale aggregation module (RMSA) to promote the salient region localization and the global-to-local context learning, respectively.
- The proposed method achieves state-of-the-art performance on six large-scale datasets including the NJUD, NLPR, RGBD135, STERE, SSD, and DUT-RGBD. Meanwhile, the model can infer at 93 FPS on a 2080Ti GPU.

2. Related work

2.1. Salient object detection

Throughout the past two decades, salient object detection has attracted the wide attention and interest of researchers. A comprehensive review of the recent progress in salient object detection, which includes both conventional methods and deep-learning-based methods, is provided in [27]. Itti et al. [28] raised the issue of saliency detection for the first time, which explore the saliency value of each pixel relative to surrounding pixels based on the center-surround contrasts and constructs a saliency map. Then the saliency maps under different contrast criteria (color, intensity, and orientation) are mixed and assembled into the final saliency map. Later, some methods [29,30] employ the local or global contrast to represent the change of appearance. Moreover, manifold ranking [31], bayesian [32], random walk [33], distance metric [34] are also utilized to calculate the saliency cues. These methods mainly focus on low-level visual features.

Recently, many CNN-based methods [35–38] have brought the performance of salient object detection to a new level. At the same time, the design of the saliency detection method no longer depends on low-level hand-crafted features, but relies on CNN to obtain the high-level semantics adaptively. Earlier methods [35,39] mainly utilize the powerful representation ability of deep features, and the deep features are directly embedded into the traditional learning model to realize the saliency detection. Wang et al. [39] employ local estimation and global search to evaluate both salient super-pixels and salient proposals. KSR [35] adopts the kernelized subspace ranking to classify all the region proposals. However, these region-based methods are time-consuming, and often fail to achieve end-to-end predictions.

To conquer the above problem, many methods begin to consider the pixel-wise saliency prediction in an end-to-end manner. Wang et al. [7] propose a multi-stage refined saliency detection method, which uses the recurrent architecture to realize the automatic error correction of the saliency prior. DSS [9] and MINet [36] generate accurate saliency map through the aggregation of features between multiple scales. SRM [8] propose a stage-wise refinement mechanism that can supplement low-level detailed features to refine the coarse saliency map progressively. Chen et al. [40] introduce a reverse attention module to guide the network focuses on the residual learning of the side-output. Zhao et al. [41] considers information interaction between decoder and encoder for the first time, and a gate function is proposed to control the flow of information between them. Kong et al. [42] present a spatial context-aware network to effectively aggregate multi-level deep features. Wu et al. [16] design a 'wider' network structure, which encodes different types of complementary features through two sub-networks with different structures, revealing saliency cues from different perspectives.

Also, some methods [37,43–46] consider boundary information to assist the saliency detection. For example, AFNet [43] proposes an attentive feedback network with a boundary-enhanced loss for precise and complete saliency detection. Wu et al. [44] propose a novel stacked cross refinement network. The framework aims to optimize salient object detection and edge detection simultaneously by stacking cross refinement unit. Wei et al. [37] propose a label decoupling framework to decompose the saliency label into body map and detail map for salient object detection. Yang et al. [46] decompose the single-channel saliency mask into a connectivity label that is more aware of structure and inter-pixel relationships, and use this label to enhance spatial consistency. Our methods introduce depth information as a supplement to assist in the implementation of saliency detection. These RGB-based methods also provide guidance for us, such as multi-scale context, stage-wise refinement.

2.2. RGB-D based salient object detection

Zhou et al. [47] provides a comprehensive review of the recent progress in RGB-D based Salient Object Detection. Early works for RGB-D based salient object Detection mainly rely on hand-crafted features, such as contrast [10,48]. In recent years, the development of CNN has greatly improved the performance of RGB-D saliency detection. Existing methods can be roughly divided into input fusion, feature fusion, and result fusion. The input fusion

strategy [17,18,49] takes the concatenated RGB and depth map as a whole input and uses a single model for processing. However, these methods directly use the low-level mixed information of the two modalities as input, it is difficult to filter out the noise in the image. Result fusion strategy [19,20] uses two separate models to process RGB and depth maps respectively, and then fuse their predicted saliency maps at the result side. Obviously, this strategy lacks the interaction between the intermediate feature information.

As a better choice, most of recent works utilize the feature fusion strategy. Some methods map the multi-level features of different modalities to the same feature space, and the better matching between multi-modal features can be realized in the new feature space. For example, PCANet [50] designs a complementarity-aware fusion module to realize the mutual supplement of information between multi-modal features. TANet [24] designs a channel-wise attention to select cross-modal complementary features. Piao et al. [51] consider the global location and local detail complementarities from two modalities, and propose a complementary interaction model to support the object location and boundary refinement. Fu et al. [52] employ a Siamese network to implement the joint learning and a densely-cooperative fusion strategy is proposed to accomplish the mutual supplement of cross-modal information. In addition, non-local [53] module is also used in cross-modal feature fusion. S^2MA [23] first computes their own modality-specific affinity matrix, then aggregates them and uses the fused matrix to update the two modalities, respectively. Fan et al. [22] present a bifurcated backbone strategy to utilize the abstract semantic information of high-level features and the finer details of low-level features reasonably, and a depth-enhanced module composed of spatial and channel attention is adopted to enhance the compatibility between multi-modal features. Huang et al. [15] consider the quality of input images during saliency prediction and utilize semantic-guided modality-weight maps to selectively filter useful feature information.

Another feature fusion-based strategy is to use the depth information as a guide. DMRA [21] proposes a depth-induced multi-scale weighting module to locate the salient region, and a new RGB-D based saliency detection dataset is proposed. PGAR [54] design a lightweight network for efficient saliency detection, in which the proposed alternate refinement strategy can avoid the destruction of the good property of RGB features when the depth map is low quality. Meantime, a multi-scale residual block is used to capture the multi-scale context of the input. Dynamic convolution is also applied to this task, Pang et al. [13] employ multi-scale dynamic filter generated by multi-modal features to guide the RGB features. It can make full use of the semantic context of the multi-modal features and enhances the representation ability of the decoder. Chen et al. [14] first use RGB information to enhance the depth map, and then use the enhanced depth map and RGB features to generate the final prediction.

Different from them, we construct a simple and effective deep interleaved backbone, which combines cross-modal features from the beginning of the input, and the fusion process runs through the whole encoder. In addition, we propose a cross-modal mutual guidance module and a residual multi-scale aggregation module to lead the fusion of multi-modal features in the decoding stage.

3. The proposed approach

In this section, we mainly introduce the deep interleaved backbone, the cross-modal mutual guidance module, and the residual multi-scale aggregation module. The overall architecture of the proposed method is illustrated in Fig. 2.

3.1. The deep interleaved backbone

One of the key points of RGB-D SOD is to combine the cross-modal information effectively. We can extract the features of RGB and depth at different levels (low, middle, high) through a two-stream CNN structure. Many methods implement the interaction between RGB and depth in the decoding stage [13,51], or simply add the depth feature to the RGB backbone in the encoding stage [22]. Different from them, we design a deep interleaved backbone, which can propagate the feature information of different modalities between the two-stream encoder. With the sequential structure of CNN, the backbone can gradually realize the mixing of multi-modal features at different levels. Specifically, we take the VGG-16 [55] to construct RGB and depth streams respectively. To avoid losing more details, we set the stride of the last block to 1. The network illustration is shown in Fig. 2. For the input RGB and depth images, each encoder can generates 5 intermediate features, which can be denoted as $\{\mathrm{E}_i^{\mathrm{r}}\}_{i=1}^{5}$ and $\{\mathrm{E}_i^{\mathrm{d}}\}_{i=1}^{5}$, respectively. The feature map E_4 and E_5 has the smallest spatial dimension, while E_1 has the largest one. In order to connect these two independent encoders to form a unified multi-modal feature encoding network, we design an information interaction module (IIM) to transmit the information between the two encoders. In particular,

$$\begin{cases} \mathrm{E}_i^1 = \mathrm{W}_1 * [\mathrm{E}_i^{\mathrm{r}}, \mathrm{E}_i^{\mathrm{d}}] + b_1, \\ \mathrm{E}_i^2 = \mathrm{W}_2 * [\mathrm{E}_i^{\mathrm{d}}, \mathrm{E}_i^{\mathrm{r}}] + b_2, \\ \widetilde{\mathrm{E}}_i^{\mathrm{r}} = \mathrm{E}_i^1 + \mathrm{E}_i^{\mathrm{r}} \\ \widetilde{\mathrm{E}}_i^{\mathrm{d}} = \mathrm{E}_i^2 + \mathrm{E}_i^{\mathrm{d}} \end{cases} \tag{1}$$

where W_i and b_i indicate the weight and bias of the convolutional layer respectively. $[\cdot,\cdot]$ is the concatenation operation of two feature maps along the channel axis. $*$ denotes the convolutional operation with a 3×3 filters. $\widetilde{\mathrm{E}}_i^{\mathrm{r}}$ and $\widetilde{\mathrm{E}}_i^{\mathrm{d}}$ represent the mixed multi-modal features. Then these updated features are fed into the next block for further feature extraction.

3.2. The cross-modal mutual guidance module

Generally speaking, the high-level semantic features generated by the encoder mainly highlight the global perception of the image, and the low-level semantic information pays more attention to the details of the image. Therefore, many previous saliency detection methods [8,40] use the deeper features of the encoder to locate the coarse salient regions, and use the shallower features as a supplement to restore the details of the salient regions. In order to better locate the salient region by combining the high-level features of different branches of the two-stream encoder, we design a cross-modal mutual guidance module to promote their interaction. Specifically, for the features $\mathrm{E}_5^{\mathrm{r}} \in \mathbb{R}^{C\times H\times W}$ and $\mathrm{E}_5^{\mathrm{d}} \in \mathbb{R}^{C\times H\times W}$, where H, W, and C represent their height, width and channel number, respectively. We first flatten them into a matrix representation with size $C \times (HW)$, then their affinity matrix can be calculated as follows:

$$\begin{aligned} &\widetilde{\mathrm{E}}_5 = \mathrm{E}_5^{\mathrm{r}} + \mathrm{E}_5^{\mathrm{d}}, \\ &\mathrm{A}_{\mathrm{r}} = \widetilde{\mathrm{E}}_5 \otimes \mathrm{E}_5^{\mathrm{r}\top}, \\ &\mathrm{A}_{\mathrm{d}} = \widetilde{\mathrm{E}}_5 \otimes \mathrm{E}_5^{\mathrm{d}\top}, \end{aligned} \tag{2}$$

where $\mathrm{A}_{\mathrm{r}} \in \mathbb{R}^{C\times C}$ and $\mathrm{A}_{\mathrm{d}} \in \mathbb{R}^{C\times C}$ represent the similarity matrixes guided by the initial mixed features $\widetilde{\mathrm{E}}_5$. $\otimes$ indicates matrix multiplication. The element $a_{i,j}$ of A represents the similarity between the i^{th} and the j^{th} spatial location in A. We further use the softmax function to normalize the similarity matrix along the second dimension:

$$\begin{aligned} &\mathrm{A}_{\mathrm{r}} = \mathrm{softmax}(\mathrm{A}_{\mathrm{r}}), \\ &\mathrm{A}_{\mathrm{d}} = \mathrm{softmax}(\mathrm{A}_{\mathrm{d}}), \end{aligned} \tag{3}$$

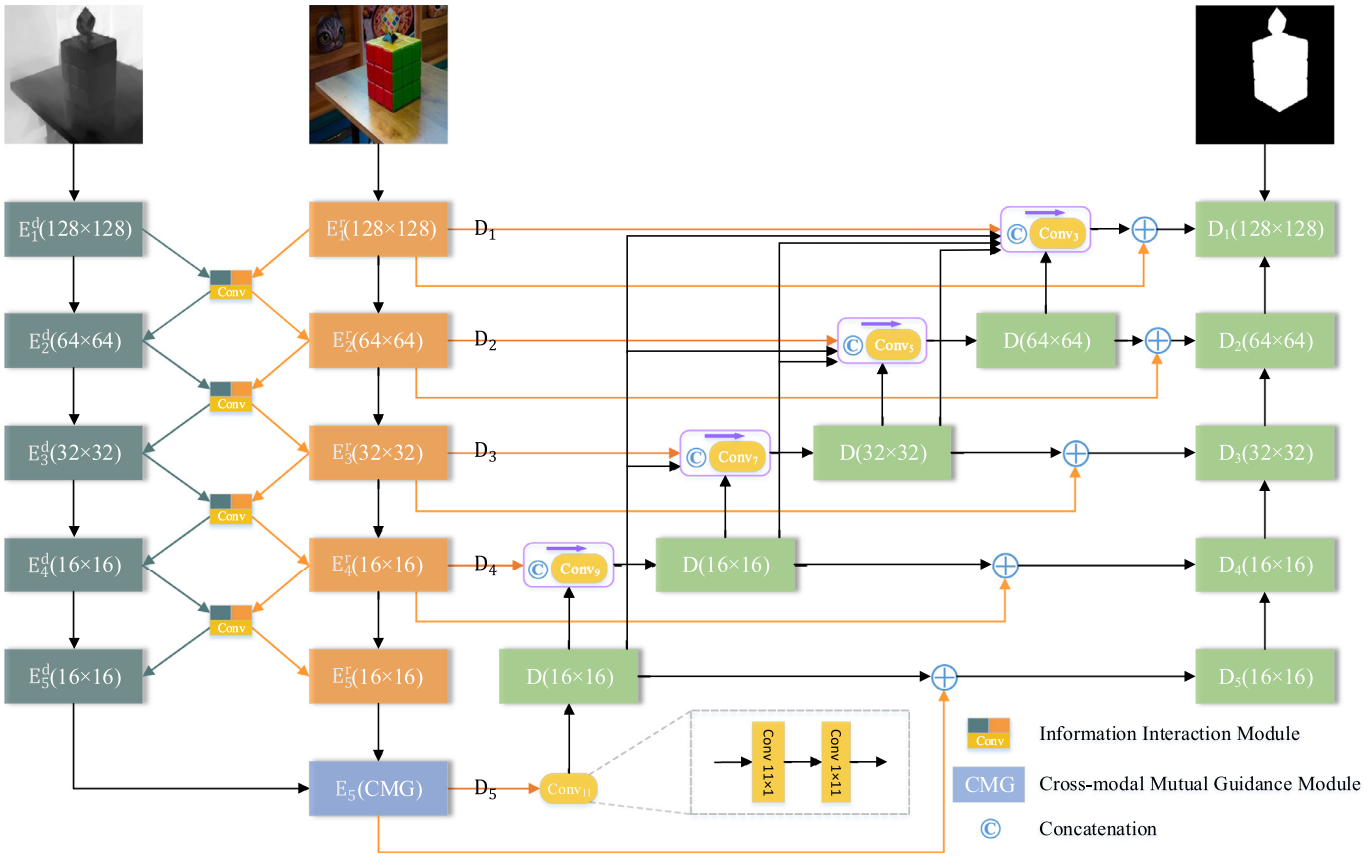


Fig. 2. The overall framework of EMANet. It consists of two VGG-16 ($E_1 \sim E_5$), a cross-modal mutual guidance module (CMG), and a residual multi-scale aggregation module (the whole decoder structure).

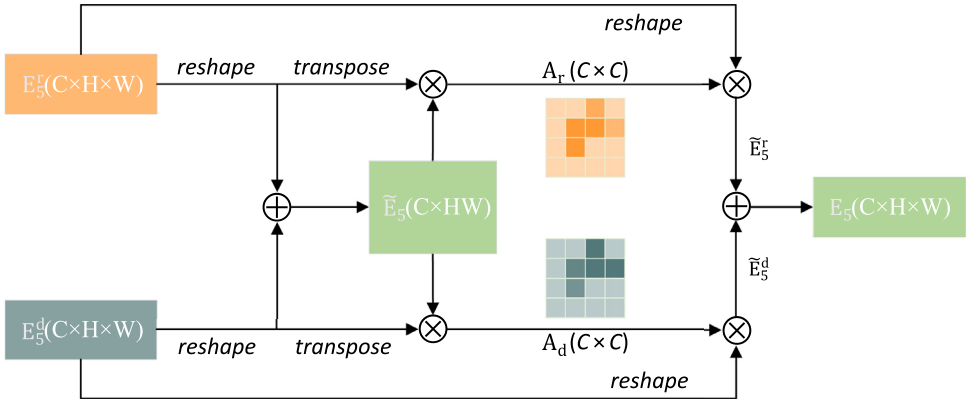


Fig. 3. The structure of cross-modal mutual guidance module.

where A_r and A_d are the results of the row-wise normalization. Subsequently, we use them to propagate the long-range contextual dependencies as follows:

$$\begin{aligned} \widetilde{E}_5^r &= A_r \otimes E_5^r, \\ \widetilde{E}_5^d &= A_d \otimes E_5^d. \end{aligned} \tag{4}$$

Finally, we add $\widetilde{E}_5^r$ and $\widetilde{E}_5^d$ to get the deepest feature representation E_5 of this two-stream encoder. Fig. 3 shows the detailed structure of the cross-modal mutual guidance module (CMG).

3.3. The residual multi-scale aggregation module

In order to effectively aggregate different levels of features and give full play to the guiding or refinement of each feature block, we introduce a residual multi-scale aggregation module (RMSA) to combine the global-to-local context progressively. The detailed structure of this module is shown in Fig. 2. At first, for the mixed multi-modal features $\{\widetilde{E}_i^r\}_{i=1}^4$ and E_5, we use 1×1 convolution to reduce the channels of each feature block of the encoder to 32 for saving memory. The features after dimensionality reduction are defined as $\{D_i\}_{i=1}^5$. The process of context aggregation can be formulated as:

$$\begin{cases} D_5 = Conv_{11}(D_5), \\ D_4 = Conv_9(Cat(D_4, D_5)), \\ D_3 = Conv_7(Cat(D_3, D_4, D_5)), \\ D_2 = Conv_5(Cat(D_2, D_3, D_4, D_5)), \\ D_1 = Conv_3(Cat(D_1, D_2, D_3, D_4, D_5)), \end{cases} \tag{5}$$

where $Cat(\cdot,\cdot)$ represents the concatenation operation. $Conv_s$ denotes the convolutional layer and the subscript s represents the size of the convolution kernel. In addition, we decompose the $s \times s$ convolution kernel into $s \times 1$ and $1 \times s$, which reduces the computational complexity while ensuring the receptive field. Subsequently, the aggregated features $\{D_i\}_{i=1}^{5}$ is recovered to the original channel number by a 1×1 convolution, and they are added to $\{\widetilde{E}_i^r\}_{i=1}^{4}$ and E_5 respectively to obtain the residual multi-scale context. RMSA uses the global context contained in high-level features as a guide to making low-level features pay more attention to the desired object geometry details. Finally, we employ the feature pyramid networks (FPN) to generate the final saliency map.

3.4. Difference to other networks

(1) Our residual multi-scale aggregation module (RMSA) is different from other top-down progressive fusion patterns, e.g. ASIF-Net [56], ICNet [57], PGAR [54], S^2MA [23]. ASIF-Net uses the saliency map of the side output to re-weight the feature of each region, and then combines the re-weighted features of each level in a top-down manner. PGAR uses the saliency map of the side output to guide the network to learn the residuals for the object region. Both of them are the deeply supervised progressive fusion network, and they have a high-quality requirement for the initial map. While ICNet and S^2MA adopt the FPN to fuse multi-level cross-modal features. Compared with the above methods, the RMSA has two advantages: 1) It does not rely on the saliency map of the side output and directly utilizes the feature maps to realize the global-to-local information guidance. We add large-size convolution to high-level features for global semantics, and add small-size convolution to low-level features for geometrical details. 2) It introduces additional skip connections for multi-feature interaction, which can capture the skip-level contextual dependencies among different scales. As a result, the RMSA sequentially aggregates the global-to-local contexts and well retains the hierarchical multi-scale information at each level.

(2) S^2MA calculates the pixel-wise affinities at the highest level, which aims to depict the local context between any pair of pixels. While our CMG calculates the channel-wise affinities, which can capture the global semantics. Since each channel of the highest-level features can be regarded as a class-specific response, by exploiting the dependencies between any two channels, different semantic responses are associated with each other, thereby promoting the overall class-independent semantic representation for saliency detection.

4. Experiments and results

4.1. Datasets

To verify the effectiveness of our proposed method, we evaluate the performance on six popular benchmark datasets: NJUD [58], NLPR [48], RGBD135 [10], STERE [59], DUT-RGBD [21], and SSD [60]. NJUD dataset contains 1985 groups of RGB, depth, and label images, which are gathered from the Internet, 3D movies, and photographs taken by a Fuji W3 stereo camera. NLPR has 1000 natural RGBD images from the Microsoft Kinect with different illumination conditions. RGBD135 contains 135 images with seven indoor scenes. STERE contains 1000 stereoscopic images downloaded from the Internet. DUT-D is a new dataset that contains 800 indoor and 400 outdoor scenes paired with the depth maps and ground truths. And it contains many complex scenes. SSD includes 80 images selected from three stereo movies.

4.2. Implementation details

The proposed network is implemented based on the public PyTorch toolbox, and it is trained on an Nvidia GTX 2080Ti GPU for 50,000 iterations. We use the SGD optimizer and set the learning rate to $6e^{-3}$ with a decay of 0.005. Pre-trained VGG-16 is used to initialize the convolutional layers in the encoder network. Since the depth image has only one channel, for the depth stream, we initialize the parameters of the first convolutional layer in block E_1^d and output a 64-channel feature. All input images are resized to 256×256 pixels for training and testing. During training, we adopt random horizontal flipping, random rotating to expand the RGB images and the depth images. Moreover, we employ random color jittering for RGB images. The batch size is set to 16.

To compare different methods fairly, we follow the setting of [13,18,21]. For DUT-RGBD, we use 800 images for training and the rest 400 for testing. For the other datasets, we select 1485 samples from the NJUD and 700 samples from the NLPR as the training set, and the remaining images and other datasets are used for testing.

4.3. Evaluation metrics

We utilize some metrics to evaluate the performance of different salient object detection algorithm, they are precision-recall (PR) curves, max F-measure (F_β^m), mean F-measure (F_β), weighted F-measure (F_β^ω) [61], Mean Absolute Error (M), S-measure (S_m) [62], and E-measure (E_m) [63]. For a given continuous saliency map, we first normalize it to [0, 255]. Then we binarize the saliency map with every possible fixed integer threshold, a sequence of precision and recall pairs are computed to plot the PR curve. The F-measure can be calculated as:

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 Precision + Recall}, \tag{6}$$

where the weight β is set to 0.3. Before calculating F-measure, we need an adaptive threshold T to binarize the saliency map. For Mean Absolute Error, it represents the difference between binary ground truth G and the predicted saliency map S:

$$\mathrm{MAE} = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} |S(i,j) - G(i,j)|, \tag{7}$$

where W and H are the width and height of an image, (i, j) denotes the location of the pixel. S-measure can be used to evaluate structure similarity in both region level and object level. This measure is based on two important characteristics: sharp foreground-background contrast and uniform saliency distribution. E-measure can represent image-level statistics and local pixel matching information.

4.4. Performance comparison

To verify the effectiveness of our model, we compare it with some state-of-the-art RGB-D based SOD methods, which are DESM [10] DCMC [11], CDCP [64], DF [49], CTMF [65], PCANet [50], MMCI [12], TANet [24], AFNet [66], CPFP [25], DMRA [21], D3Net [26], DANet [18], CoNet [67], and BBS-Net [22].

Performance Evaluation. Table 1 and Table 2 show the performance comparison of different methods on six datasets. Fig. 4 and Fig. 5 illustrate the PR curves and the F-measure curves respectively. We can observe from these results that the proposed method consistently outperforms these competitors on most datasets. Especially on the STERE, comparing F_β, F_β^ω, and E_m, our model outperforms the second-best method by 2.2%, 2.6% 1.2%. And the MAE value is 13% lower than the second-best. STERE contains 1000 test images, which is a relatively large RGB-D dataset

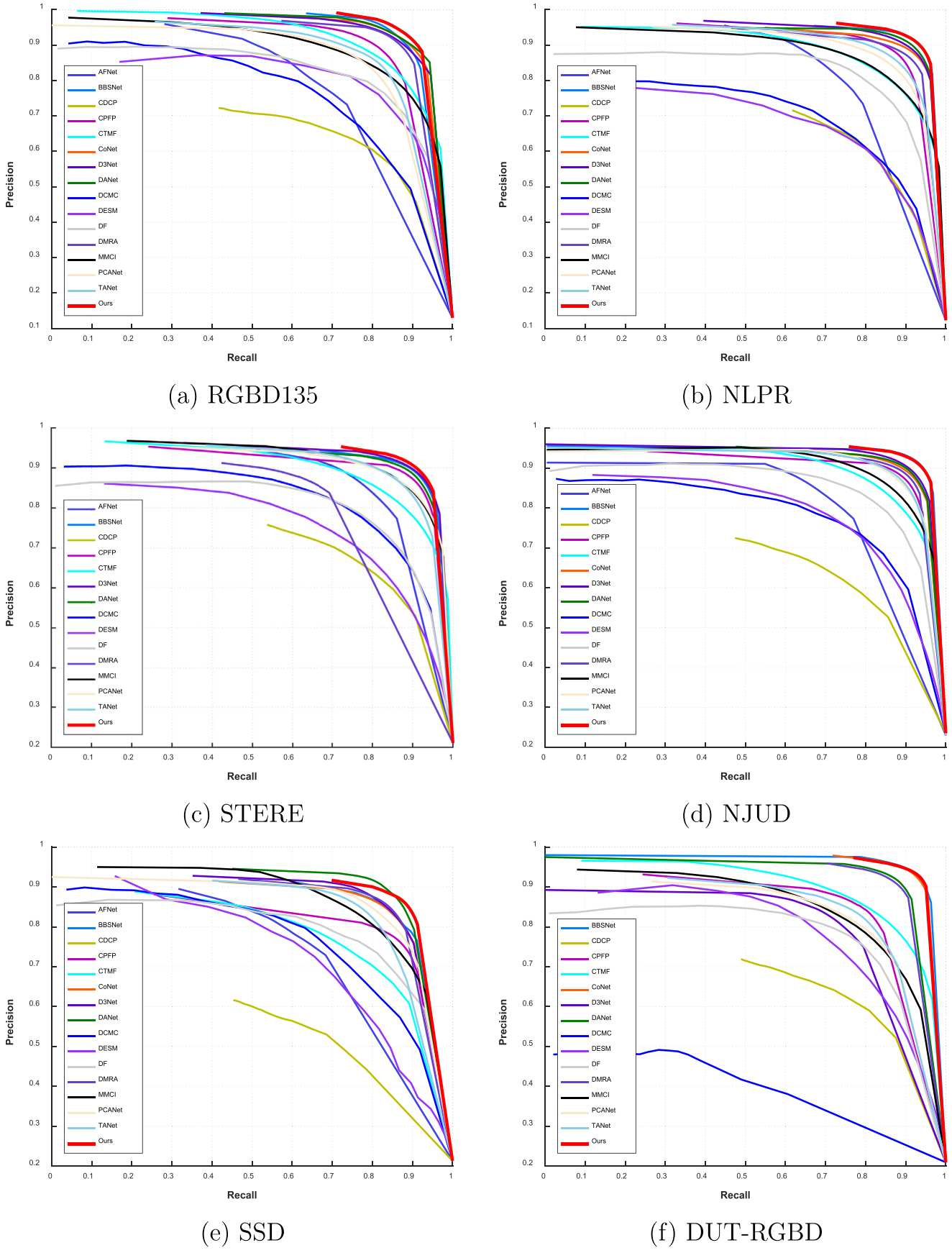
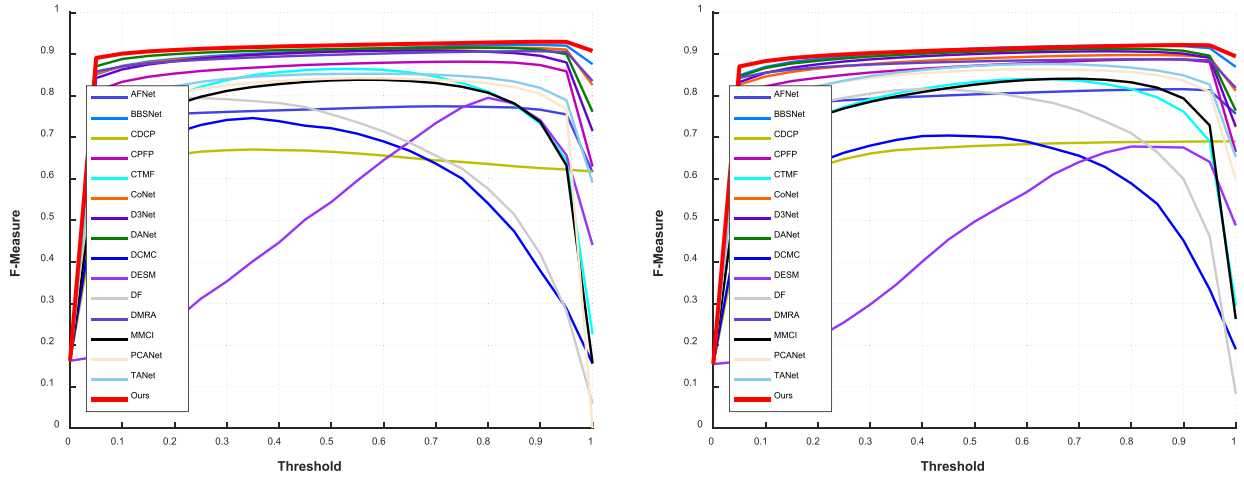
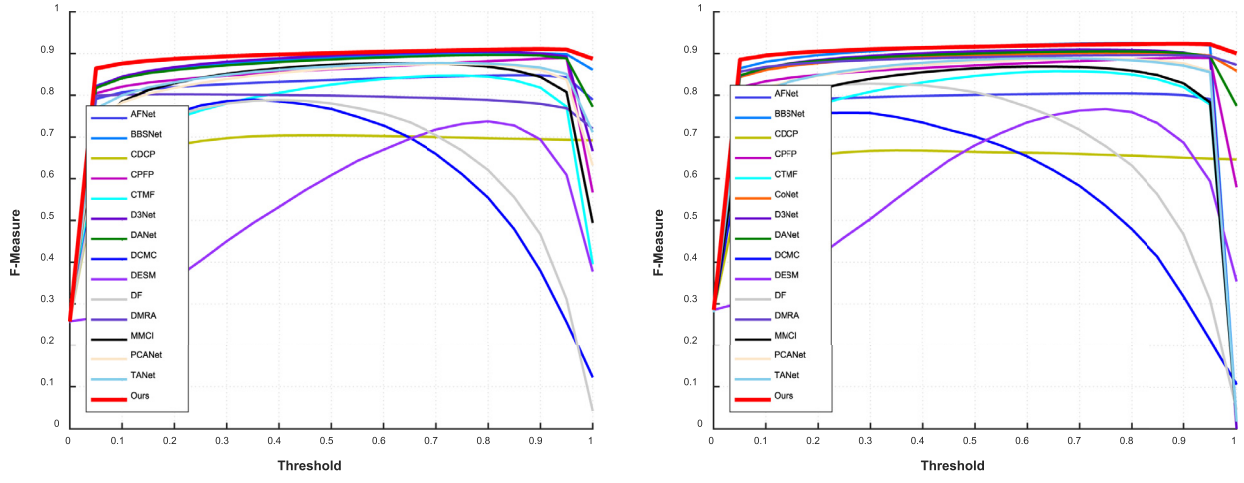


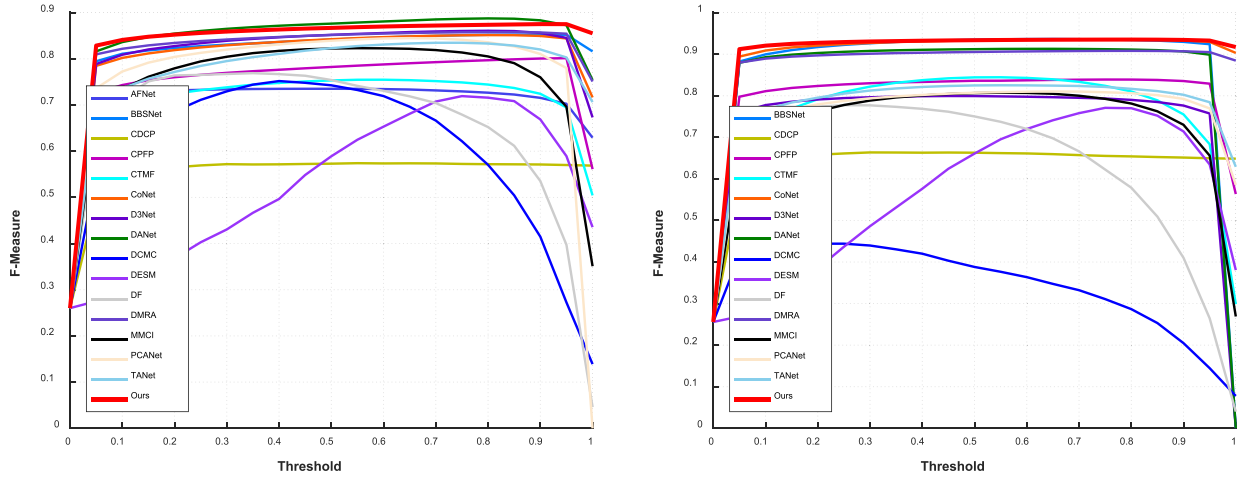
Fig. 4. The PR curves of proposed approach and other baseline methods on six datasets.



(a) RGBD135 (b) NLPR



(c) STERE (d) NJUD



(e) SSD (f) DUT-RGBD

Fig. 5. The F-measure curves of proposed approach and other baseline methods on six datasets.

Table 1
Quantitative evaluation in terms of F_β^m, F_β, F_β^ω, M, S_m, and E_m. the best results are shown in bold. Where the subscript represents the year the paper was published.

*	RGBD135						NLPR						STERE					
	$F_\beta^m \uparrow$	$F_\beta \uparrow$	$F_\beta^\omega \uparrow$	M $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^m \uparrow$	$F_\beta \uparrow$	$F_\beta^\omega \uparrow$	M $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^m \uparrow$	$F_\beta \uparrow$	$F_\beta^\omega \uparrow$	M $\downarrow$	$S_m \uparrow$	$E_m \uparrow$
Our	**0.930**	**0.897**	**0.873**	**0.023**	**0.915**	**0.953**	**0.922**	**0.885**	**0.882**	**0.024**	**0.924**	**0.955**	**0.911**	**0.883**	**0.860**	**0.040**	**0.901**	**0.939**
$BBSNet_{20}$	0.923	0.869	0.845	0.028	0.908	0.941	0.921	0.873	0.871	0.026	0.923	0.948	0.901	0.864	0.838	0.046	0.896	0.928
$CoNet_{20}$	0.914	0.871	0.849	0.028	0.909	0.945	0.897	0.848	0.843	0.031	0.907	0.934	-	-	-	-	-	-
$DANet_{20}$	0.916	0.891	0.848	0.028	0.905	0.961	0.913	0.871	0.858	0.028	0.915	0.949	0.897	0.858	0.830	0.047	0.892	0.926
$D3Net_{20}$	0.909	0.870	0.829	0.031	0.898	0.951	0.907	0.861	0.849	0.030	0.912	0.945	0.904	0.859	0.838	0.046	0.899	0.924
$DMRA_{19}$	0.906	0.867	0.843	0.030	0.899	0.944	0.888	0.855	0.839	0.031	0.898	0.942	0.802	0.762	0.647	0.087	0.752	0.816
$CPFP_{19}$	0.882	0.829	0.787	0.038	0.872	0.927	0.888	0.823	0.813	0.036	0.888	0.924	0.889	0.830	0.817	0.051	0.879	0.907
$AFNet_{19}$	0.775	0.730	0.641	0.068	0.770	0.874	0.816	0.747	0.693	0.058	0.799	0.884	0.848	0.807	0.752	0.075	0.825	0.887
$TANet_{19}$	0.853	0.795	0.739	0.046	0.858	0.919	0.877	0.796	0.780	0.041	0.886	0.916	0.878	0.835	0.787	0.060	0.871	0.916
$MMCI_{19}$	0.839	0.762	0.650	0.065	0.848	0.904	0.841	0.730	0.676	0.059	0.856	0.872	0.877	0.829	0.760	0.068	0.873	0.905
$PCANet_{18}$	0.842	0.774	0.711	0.050	0.843	0.912	0.864	0.795	0.762	0.044	0.873	0.916	0.875	0.826	0.778	0.064	0.875	0.907
$CTMF_{17}$	0.865	0.778	0.686	0.055	0.863	0.911	0.841	0.724	0.679	0.056	0.860	0.869	0.848	0.771	0.698	0.086	0.848	0.870
DF_{17}	0.796	0.753	0.518	0.093	0.752	0.877	0.817	0.759	0.592	0.079	0.806	0.884	0.789	0.742	0.549	0.141	0.757	0.838
$CDCP_{17}$	0.671	0.625	0.486	0.115	0.709	0.816	0.690	0.608	0.512	0.108	0.732	0.804	0.704	0.666	0.558	0.149	0.713	0.796
$DCMC_{16}$	0.747	0.702	0.445	0.111	0.707	0.849	0.706	0.603	0.458	0.112	0.729	0.778	0.789	0.742	0.520	0.148	0.731	0.831
$DESM_{14}$	0.802	0.698	0.296	0.299	0.622	0.795	0.680	0.575	0.250	0.309	0.573	0.749	0.738	0.594	0.375	0.295	0.642	0.696

Table 2
Quantitative evaluation in terms of F_β^m, F_β, F_β^ω, M, S_m, and E_m. the best results are shown in bold. Where the subscript represents the year the paper was published.

*	NJUD						SSD						DUT-RGBD					
	$F_\beta^m \uparrow$	$F_\beta \uparrow$	$F_\beta^\omega \uparrow$	M $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^m \uparrow$	$F_\beta \uparrow$	$F_\beta^\omega \uparrow$	M $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^m \uparrow$	$F_\beta \uparrow$	$F_\beta^\omega \uparrow$	M $\downarrow$	$S_m \uparrow$	$E_m \uparrow$
Our	0.923	**0.903**	**0.888**	**0.035**	0.914	**0.946**	0.875	**0.837**	**0.809**	**0.047**	**0.870**	**0.909**	0.937	**0.914**	**0.900**	**0.032**	**0.920**	0.951
$BBSNet_{20}$	**0.924**	0.894	0.881	0.039	**0.915**	0.936	0.852	0.815	0.774	0.055	0.857	0.901	0.938	0.908	0.888	0.037	**0.920**	0.949
$CoNet_{20}$	0.902	0.873	0.856	0.046	0.895	0.924	0.851	0.806	0.781	0.060	0.852	0.898	0.936	0.909	0.896	0.033	0.919	**0.952**
$DANet_{20}$	0.905	0.877	0.853	0.046	0.897	0.926	**0.888**	0.831	0.798	0.050	0.869	**0.909**	0.913	0.884	0.852	0.047	0.889	0.929
$D3Net_{20}$	0.903	0.840	0.833	0.051	0.895	0.901	0.861	0.814	0.777	0.058	0.857	0.904	0.796	0.756	0.675	0.097	0.775	0.849
$DMRA_{19}$	0.896	0.872	0.847	0.051	0.885	0.920	0.858	0.821	0.787	0.058	0.856	0.898	0.908	0.883	0.857	0.048	0.888	0.930
$CPFP_{19}$	0.890	0.837	0.828	0.053	0.878	0.900	0.801	0.726	0.708	0.082	0.807	0.832	0.837	0.794	0.743	0.076	0.818	0.869
$AFNet_{19}$	0.804	0.766	0.697	0.098	0.774	0.853	0.735	0.694	0.589	0.118	0.714	0.803	-	-	-	-	-	-
$TANet_{19}$	0.882	0.838	0.797	0.063	0.874	0.909	0.835	0.767	0.726	0.063	0.839	0.886	0.823	0.779	0.712	0.093	0.808	0.871
$MMCI_{19}$	0.868	0.813	0.739	0.079	0.859	0.882	0.823	0.748	0.660	0.082	0.813	0.860	0.804	0.753	0.636	0.112	0.791	0.856
$PCANet_{18}$	0.887	0.844	0.803	0.059	0.877	0.909	0.844	0.786	0.733	0.063	0.842	0.890	0.809	0.760	0.696	0.100	0.801	0.863
$CTMF_{17}$	0.857	0.788	0.720	0.085	0.849	0.866	0.755	0.710	0.624	0.099	0.776	0.839	0.842	0.792	0.690	0.097	0.831	0.882
DF_{17}	0.789	0.744	0.545	0.151	0.735	0.818	0.769	0.724	0.549	0.142	0.747	0.812	0.775	0.748	0.542	0.145	0.730	0.842
$CDCP_{17}$	0.661	0.618	0.510	0.182	0.672	0.751	0.574	0.522	0.423	0.214	0.603	0.705	0.659	0.633	0.530	0.159	0.687	0.794
$DCMC_{16}$	0.769	0.715	0.497	0.167	0.703	0.796	0.755	0.679	0.477	0.169	0.704	0.786	0.444	0.406	0.290	0.243	0.499	0.712
$DESM_{14}$	0.328	0.165	0.234	0.448	0.413	0.491	0.720	0.614	0.343	0.308	0.602	0.704	0.771	0.668	0.386	0.280	0.659	0.751

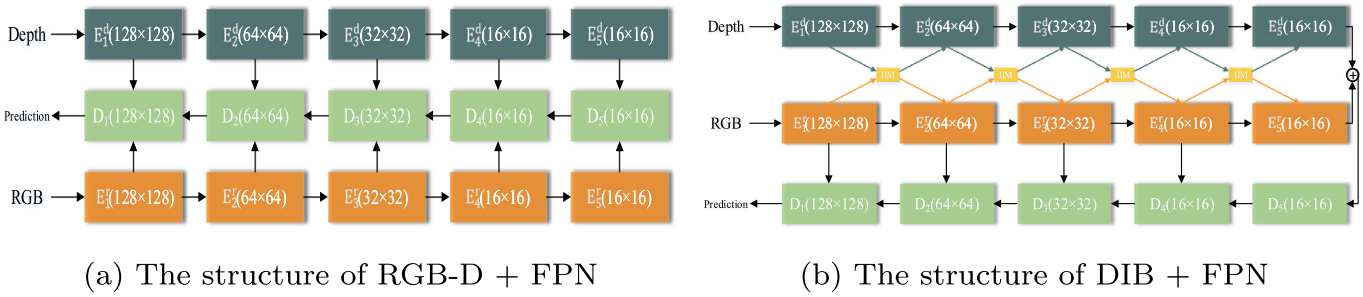


Fig. 6. Various baselines. IIM represents the information interaction module.

Fig. 7. Visual comparison with state-of-the-art methods.

at present. This proves that our algorithm is more stable on large-scale datasets.

To qualitatively evaluate the proposed method, we also give some visual examples in Fig. 7. In the first two columns, compared with previous methods, our method can segment the narrow part (e.g. foliage) of objects more clearly even with unclear depth map. In the 3rd column, the foreground and background contain very similar information. Other methods can not suppress the background regions accurately with the help of the depth map, while our method can segment the entire foreground region accurately. The boundaries of the low-contrast region are usually difficult to be extracted (See the 4th to 8th columns). Some previous methods incorrectly distinguish the foreground and background of the low-contrast region. However, the proposed method performs well. In the 10th column, the depth map provides ambiguous guidance information, which can mislead the results of saliency detection. However, compared with other methods, our method can still highlight the most salient objects. Besides, in the 5th column, we also show the case of small objects, our method produces a more refined saliency map. Overall, for these challenging examples, we predict more accurate and complete saliency maps. Moreover, during the inference stage, the network is in a fully convolutional fashion running at about 93 FPS and does not need any post-processing.

4.5. Ablation study

We show a detailed ablation study in Table 3 to verify the benefit of each component. RGB means that only the RGB-based backbone is used. RGB-D indicates that only the highest-level features (E_5^d and E_5^r) are fused in the RGB and depth encoders. DIB represents the deep interleaved backbone. FPN is the feature pyra-

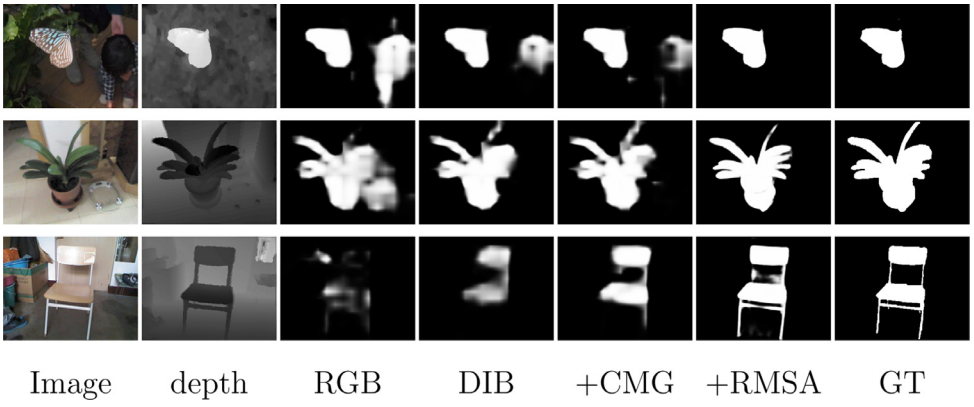


Fig. 8. Visual effect of the proposed modules.

Table 3
Ablation study on the NJUD and RGBD135 datasets..

*	RGB	RGB-D	DIB	CMG	FPN	RMSA	NJUD						RGBD135					
							$F_\beta^m \uparrow$	$F_\beta \uparrow$	$F_\beta^\omega \uparrow$	M $\downarrow$	$S_m \uparrow$	$E_m \uparrow$	$F_\beta^m \uparrow$	$F_\beta \uparrow$	$F_\beta^\omega \uparrow$	M $\downarrow$	$S_m \uparrow$	$E_m \uparrow$
	✓						0.871	0.813	0.783	0.068	0.865	0.886	0.885	0.817	0.774	0.041	0.883	0.925
		✓					0.878	0.832	0.797	0.064	0.874	0.898	0.891	0.830	0.784	0.039	0.882	0.933
		✓			✓		0.901	0.855	0.841	0.051	0.892	0.914	0.905	0.848	0.813	0.031	0.892	0.937
			✓		✓		0.909	0.875	0.862	0.043	0.897	0.923	0.911	0.888	0.832	0.029	0.892	0.939
*			✓				0.893	0.850	0.820	0.057	0.886	0.908	0.907	0.851	0.807	0.034	0.891	0.940
			✓	✓			0.899	0.854	0.833	0.053	0.891	0.913	0.908	0.856	0.822	0.033	0.900	0.939
			✓	✓	✓		0.913	0.884	0.868	0.041	0.905	0.932	0.922	0.884	0.857	0.027	0.910	0.946
			✓	✓		✓	0.923	0.903	0.888	0.035	0.914	0.946	0.930	0.897	0.873	0.023	0.915	0.953

mid network. The structures of 'RGB-D + FPN' and 'DIB + FPN' are shown in Fig. 6. CMG denotes the cross-modal mutual guidance module. RMSA stands for the residual multi-scale aggregation module. Comparing the first two rows, the results demonstrate that the depth map is beneficial for saliency detection. The comparison between 'RGB-D' and 'RGB-D + FPN' proves that the fusion of multi-level cross-modal features can further improve the performance. DIB is a multi-level cross-modal fusion structure. To verify its advantages fairly, we compare the results of 'RGB-D + FPN' and 'DIB + FPN'. The difference between them is that they employ different encoders (as shown in Fig. 6). 'DIB + FPN' shows obvious performance advantages under all six metrics. In addition, we add CMG to the DIB to further capture the whole shape of salient region. The comparison between the last two rows shows that RMSA can harvest rich context information to help the network to detect and segment salient region more accurately. In Fig. 8, we also show some visualization results.

Table 4
Run time cost of adding DIB, CMG, and RMSA respectively.

*	Base	+DIB	+CMG	+RMSA
Time (ms)	5.66	8.70	9.09	10.71

Runtime Analysis. We also demonstrate the running time of different modules in Table 4. All the tests are implemented on an NVIDIA RTX 2080 Ti GPU.

4.6. Failure cases

We give some failure cases as shown in Fig. 9. The cluttered backgrounds possibly result in false detection and the transparent objects may lead to incomplete segmentation. In the future, we can

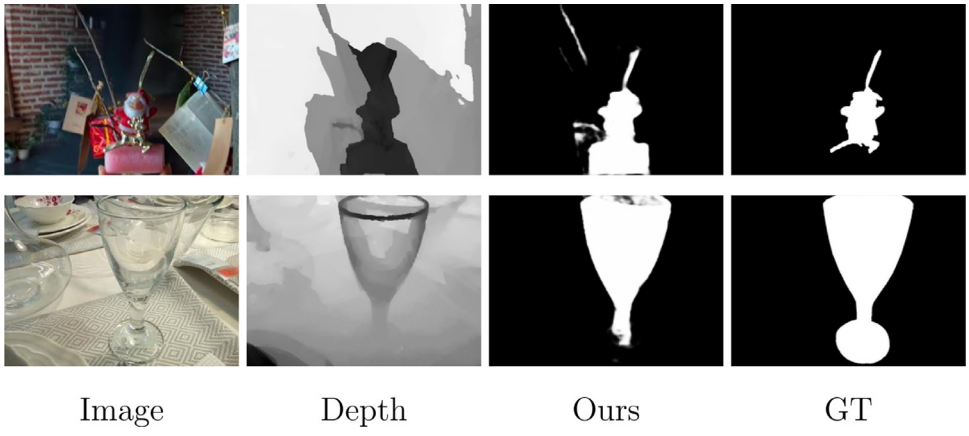


Fig. 9. Failure cases for objects with low-contrast scenes and very cluttered backgrounds.

introduce a boundary-aware mechanism to capture more complete and clearer salient region.

5. Conclusions

In this paper, we propose an end-to-end deep neural network for RGB-D salient object detection. To obtain multi-level continuous multi-modal features, a unified two-stream framework is designed to couple the RGB features and depth features progressively. This new framework can directly obtain the fused features of the two inputs of RGB image and depth image, and we do not need to design the interaction process of cross-modal features carefully. Then, a cross-modal mutual guidance module and a residual multi-scale aggregation module is proposed to assist the location and segmentation of the salient region. Finally, extensive evaluations demonstrate that our method outperforms previous state-of-the-art methods. And our method runs at a real-time speed. In the future, we can introduce the quality evaluation of the depth map at the input end. Since the IIM utilizes the 'addition' operation to fuse multi-modal features, for the low-quality depth maps, it can directly mask the depth branch at the encoding end without affecting the forward propagation of the network and only rely on the RGB image to predict the salient region.

Declaration of Competing 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, Encoder Deep Interleaved Network with Multi-scale Aggregation for RGB-D Salient Object Detection

Acknowledgements

This work was supported by the National Natural Science Foundation of China #61876202, the Liaoning Province Natural Science Foundation #2021-KF-12-10, and the Fundamental Research Funds for the Central Universities #DUT20ZD212.

References

[1] S. Chen, Q. Zhao, Boosted attention: leveraging human attention for image captioning, in: Proceedings of European Conference on Computer Vision, 2018, pp. 68–84.
[2] Y. Gao, M. Wang, Z.-J. Zha, J. Shen, X. Li, X. Wu, Visual-textual joint relevance learning for tag-based social image search, IEEE Trans. Image Process. 22 (1) (2012) 363–376.
[3] Y. Wei, X. Liang, Y. Chen, X. Shen, M.-M. Cheng, J. Feng, Y. Zhao, S. Yan, Stc: a simple to complex framework for weakly-supervised semantic segmentation, IEEE Trans Pattern Anal Mach Intell 39 (11) (2016) 2314–2320.
[4] M.-M. Cheng, Y. Liu, W.-Y. Lin, Z. Zhang, P.L. Rosin, P.H.S. Torr, Bing: binarized normed gradients for objectness estimation at 300fps, Computational Visual Media 5 (1) (2019) 3–20.
[5] S. Hong, T. You, S. Kwak, B. Han, Online tracking by learning discriminative saliency map with convolutional neural network, in: Proceedings of International Conference on Machine Learning, 2015, pp. 597–606.
[6] A.W. Smeulders, D.M. Chu, R. Cucchiara, S. Calderara, A. Dehghan, M. Shah, Visual tracking: an experimental survey, IEEE Trans Pattern Anal Mach Intell 36 (7) (2013) 1442–1468.
[7] L. Wang, L. Wang, H. Lu, P. Zhang, X. Ruan, Saliency detection with recurrent fully convolutional networks, in: Proceedings of European Conference on Computer Vision, Springer, 2016, pp. 825–841.
[8] T. Wang, A. Borji, L. Zhang, P. Zhang, H. Lu, A stagewise refinement model for detecting salient objects in images, in: Proceedings of IEEE International Conference on Computer Vision, 2017, pp. 4019–4028.
[9] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, P.H. Torr, Deeply supervised salient object detection with short connections, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3203–3212.
[10] Y. Cheng, H. Fu, X. Wei, J. Xiao, X. Cao, Depth enhanced saliency detection method, in: Proceedings of International Conference on Internet Multimedia Computing and Service, 2014, pp. 23–27.
[11] R. Cong, J. Lei, C. Zhang, Q. Huang, X. Cao, C. Hou, Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion, IEEE Signal Process Lett 23 (6) (2016) 819–823.
[12] H. Chen, Y. Li, D. Su, Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection, Pattern Recognit 86 (2019) 376–385.
[13] Y. Pang, L. Zhang, X. Zhao, H. Lu, Hierarchical dynamic filtering network for rgb-d salient object detection, in: European Conference on Computer Vision, Springer, 2020, pp. 235–252.
[14] Q. Chen, K. Fu, Z. Liu, G. Chen, H. Du, B. Qiu, L. Shao, Ef-net: a novel enhancement and fusion network for rgb-d saliency detection, Pattern Recognit 112 (2021) 107740.
[15] N. Huang, Y. Luo, Q. Zhang, J. Han, Discriminative unimodal feature selection and fusion for rgb-d salient object detection, Pattern Recognit 122 (2022) 108359.
[16] Z. Wu, S. Li, C. Chen, A. Hao, H. Qin, Recursive multi-model complementary deep fusion for robust salient object detection via parallel sub-networks, Pattern Recognit 121 (2022) 108212.
[17] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, T. Ren, Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning, IEEE Trans. Image Process. 26 (9) (2017) 4204–4216.
[18] X. Zhao, L. Zhang, Y. Pang, H. Lu, L. Zhang, A single stream network for robust and real-time rgb-d salient object detection (2020) arXiv preprint arXiv:2007.06811.
[19] X. Fan, Z. Liu, G. Sun, Salient region detection for stereoscopic images, in: 2014 19th International Conference on Digital Signal Processing, IEEE, 2014, pp. 454–458.
[20] J. Guo, T. Ren, J. Bei, Salient object detection for rgb-d image via saliency evolution, in: 2016 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2016, pp. 1–6.
[21] Y. Piao, W. Ji, J. Li, M. Zhang, H. Lu, Depth-induced multi-scale recurrent attention network for saliency detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7254–7263.
[22] D.-P. Fan, Y. Zhai, A. Borji, J. Yang, L. Shao, Bbs-net: Rgb-d salient object detection with a bifurcated backbone strategy network, in: European Conference on Computer Vision, Springer, 2020, pp. 275–292.
[23] N. Liu, N. Zhang, J. Han, Learning selective self-mutual attention for rgb-d saliency detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13756–13765.
[24] H. Chen, Y. Li, Three-stream attention-aware network for RGB-D salient object detection, IEEE Trans. Image Process. 28 (6) (2019) 2825–2835.
[25] J.-X. Zhao, Y. Cao, D.-P. Fan, M.-M. Cheng, X.-Y. Li, L. Zhang, Contrast prior and fluid pyramid integration for rgbd salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3927–3936.
[26] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, M.-M. Cheng, Rethinking rgb-d salient object detection: models, data sets, and large-scale benchmarks, IEEE Trans Neural Netw Learn Syst (2020).
[27] A. Borji, M.-M. Cheng, Q. Hou, H. Jiang, J. Li, Salient object detection: a survey, Computational visual media (2019) 1–34.
[28] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans Pattern Anal Mach Intell 20 (11) (1998) 1254–1259.
[29] S. Goferman, L. Zelnik-Manor, A. Tal, Context-aware saliency detection, IEEE Trans Pattern Anal Mach Intell 34 (10) (2011) 1915–1926.
[30] M.-M. Cheng, N.J. Mitra, X. Huang, P.H. Torr, S.-M. Hu, Global contrast based salient region detection, IEEE Trans Pattern Anal Mach Intell 37 (3) (2014) 569–582.
[31] L. Zhang, C. Yang, H. Lu, X. Ruan, M.-H. Yang, Ranking saliency, IEEE Trans Pattern Anal Mach Intell 39 (9) (2016) 1892–1904.
[32] X. Li, H. Lu, L. Zhang, X. Ruan, M.-H. Yang, Saliency detection via dense and sparse reconstruction, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 2976–2983.
[33] B. Jiang, L. Zhang, H. Lu, C. Yang, M.-H. Yang, Saliency detection via absorbing Markov Chain, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1665–1672.
[34] D.A. Klein, S. Frintrop, Center-surround divergence of feature statistics for salient object detection, in: 2011 International Conference on Computer Vision, IEEE, 2011, pp. 2214–2219.
[35] T. Wang, L. Zhang, H. Lu, C. Sun, J. Qi, Kernelized subspace ranking for saliency detection, in: European Conference on Computer Vision, Springer, 2016, pp. 450–466.
[36] Y. Pang, X. Zhao, L. Zhang, H. Lu, Multi-scale interactive network for salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9413–9422.
[37] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, Q. Tian, Label decoupling framework for salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 13025–13034.
[38] Y. Lv, B. Liu, J. Zhang, Y. Dai, A. Li, T. Zhang, Semi-supervised active salient object detection, Pattern Recognit 123 (2022) 108364.
[39] L. Wang, H. Lu, X. Ruan, M.-H. Yang, Deep networks for saliency detection via local estimation and global search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3183–3192.
[40] S. Chen, X. Tan, B. Wang, X. Hu, Reverse attention for salient object detection, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 234–250.

[41] X. Zhao, Y. Pang, L. Zhang, H. Lu, L. Zhang, Suppress and balance: a simple gated network for salient object detection (2020) arXiv preprint arXiv:2007.08074..
[42] Y. Kong, M. Feng, X. Li, H. Lu, X. Liu, B. Yin, Spatial context-aware network for salient object detection, Pattern Recognit 114 (2021) 107867.
[43] M. Feng, H. Lu, E. Ding, Attentive feedback network for boundary-aware salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1623–1632.
[44] Z. Wu, L. Su, Q. Huang, Stacked cross refinement network for edge-aware salient object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7264–7273.
[45] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, E. Ding, A mutual learning method for salient object detection with intertwined multi-supervision, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 8150–8159.
[46] Z. Yang, S. Soltanian-Zadeh, S. Farsiu, Biconnet: an edge-preserved connectivity-based approach for salient object detection, Pattern Recognit 121 (2022) 108231.
[47] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, L. Shao, Rgb-d salient object detection: a survey, Computational Visual Media (2021) 1–33.
[48] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, RGBD salient object detection: a benchmark and algorithms, in: European Conference on Computer Vision, Springer, 2014, pp. 92–109.
[49] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, Q. Yang, Rgbd salient object detection via deep fusion, IEEE Trans. Image Process. 26 (5) (2017) 2274–2285.
[50] H. Chen, Y. Li, Progressively complementarity-aware fusion network for RGB-D salient object detection, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3051–3060.
[51] M. Zhang, W. Ren, Y. Piao, Z. Rong, H. Lu, Select, supplement and focus for RGB-D saliency detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3472–3481.
[52] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, JL-DCF: Joint learning and densely-cooperative fusion framework for rgb-d salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 3052–3062.
[53] X. Wang, R. Girshick, A. Gupta, K. He, Non-local neural networks, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 7794–7803.
[54] S. Chen, Y. Fu, Progressively guided alternate refinement network for rgb-d salient object detection, in: European Conference on Computer Vision, Springer, 2020, pp. 520–538.
[55] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition (2014) arXiv preprint arXiv:1409.1556..
[56] C. Li, R. Cong, S. Kwong, J. Hou, H. Fu, G. Zhu, D. Zhang, Q. Huang, Asif-net: attention steered interweave fusion network for RGB-D salient object detection, IEEE Trans Cybern 51 (1) (2020) 88–100.
[57] G. Li, Z. Liu, H. Ling, Icnet: information conversion network for rgb-d based salient object detection, IEEE Trans. Image Process. 29 (2020) 4873–4884.
[58] R. Ju, L. Ge, W. Geng, T. Ren, G. Wu, Depth saliency based on anisotropic center-surround difference, in: 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 1115–1119.
[59] Y. Niu, Y. Geng, X. Li, F. Liu, Leveraging stereopsis for saliency analysis, in: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 454–461.
[60] C. Zhu, G. Li, A three-pathway psychobiological framework of salient object detection using stereoscopic technology, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 3008–3014.
[61] R. Margolin, L. Zelnik-Manor, A. Tal, How to evaluate foreground maps? in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 248–255.
[62] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, A. Borji, Structure-measure: A new way to evaluate foreground maps, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4548–4557.
[63] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, A. Borji, Enhanced-alignment measure for binary foreground map evaluation (2018) arXiv preprint arXiv:1805.10421..
[64] C. Zhu, G. Li, W. Wang, R. Wang, An innovative salient object detection using center-dark channel prior, in: ICCV Workshop, 2017, pp. 1509–1515.
[65] J. Han, H. Chen, N. Liu, C. Yan, X. Li, Cnns-based RGB-D saliency detection via cross-view transfer and multiview fusion, IEEE Trans Cybern 48 (11) (2017) 3171–3183.
[66] N. Wang, X. Gong, Adaptive fusion for RGB-D salient object detection, IEEE Access 7 (2019) 55277–55284.
[67] W. Ji, J. Li, M. Zhang, Y. Piao, H. Lu, Accurate RGB-D salient object detection via collaborative learning, in: European Conference on Computer Vision, Springer, 2020, pp. 52–69.

Guang Feng received his B.E. degree in electronic information engineering from Qingdao University, Qingdao, China and M.E. degree in signal and information processing from the University of Jinan, Jinan, China in 2015 and 2018 respectively. He is currently a Ph.D. candidate in the School of Information and Communication Engineering, Dalian University of Technology, Dalian, China. His research interests include saliency detection and referring expression comprehension.

Jinyu Meng received his B.E. degree in electronic information engineering from Dalian University of Technology, Dalian, China, in 2020. He is currently pursuing the M.S. degree in the School of Information and Communication Engineering, Dalian University of Technology, Dalian, China. His research interest is referring expression comprehension.

Lihe Zhang received the M.S. degree and the Ph.D. degree in Signal and Information Processing from Harbin Engineering University (HEU), Harbin, China, in 2001 and from Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2004, respectively. He is currently an Full Professor with the School of Information and Communication Engineering, Dalian University of Technology (DUT). His current research interests include computer vision and pattern recognition.

Huchuan Lu received the Ph.D. degree in System Engineering and the M.S. degree in Signal and Information Processing from Dalian University of Technology (DUT), Dalian, China, in 2008 and 1998, respectively. He joined the faculty in 1998 and currently is a Full Professor of the School of Information and Communication Engineering, DUT. His current research interests include computer vision and pattern recognition with focus on visual tracking, saliency detection, and segmentation. He is a member of the ACM and an Associate Editor of the IEEE Transactions on Cybernetics.