ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 17, No. 2, 2022, *pp*.102-117

LTDNet: A lightweight two-stage decoder network for RGB-D salient object detection

Jian Wang, Wenbing Chen[1]

School of Mathematics and Statistics, Nanjing University of Information Science & Technology, Nanjing 210044, China

(Received August 17, 2022, accepted October 20, 2022)

Abstract: Most existing models of RGB-D salient object detection (SOD) utilize heavy backbones like VGGs and ResNets which lead to large model size and high computational costs. In order to improve this problem, a lightweight two-stage decoder network is proposed. Firstly, the network utilizes MobileNet-V2 and a customized backbone to extract the features of RGB images and depth maps respectively. In order to mine and combine cross-modality information, cross reference module is used to fuse complementary information from different modalities. Subsequently, we design a feature enhancement module to enhance the clues of the fused features which has four parallel convolutions with different expansion rates. Finally, a two-stage decoder is used to predict the saliency maps, which processes high-level features and low-level features separately and then merges them. Experiments on 5 benchmark datasets comparing with 10 state-of-the-art models demonstrate that our model can achieve significant improvement with smallest model size.

Keywords: salient object detection, RGB-D, lightweight, efficient

1. Introduction

Salient object detection (SOD) aims to locate and segment the most eye-catching objects in a scene by simulating the human visual attention mechanism. SOD has been developed rapidly due to its wide application in image processing and computer vision, such as visual tracking [1], image segmentation [2], face recognition [3], medical segmentation [4] and so on. In the past years, the development of deep learning has driven SOD to achieve promising performance. Most existing SOD methods focus RGB images. However, it is difficult to get outstanding result in complex senses, such as camouflaged objects, similar texture, complex backgrounds, transparent objects, low-contrast.

With the popularity of depth device, depth sensor has been widely introduced into different fields to capture depth maps, which can provide additional clues for RGB images, such as object edges, 3D distribution, spatial structure. Many recent works [5-9] have been proposed and demonstrated that it is effective to improve efficiency and performance using auxiliary depth maps to assist RGB images for SOD. Although RGB-D SOD has achieved extraordinary results [10-16] in recent years, most methods use cumbersome networks as backbones which bring large model size and high computational costs, such as Resnets, VGGs. This makes it difficult to apply these methods to the devices with poor computing power.

In this paper, we propose a lightweight two-stage decoder network (LTDNet) for RGB-D SOD, which possesses smaller size and lower computational costs. We employ MobileNet-V2 to extract the features of RGB, which reduces the computational cost and network size significantly. For depth stream backbone, we

[1] Corresponding author. *E-mail address*: 001101@nuist.edu.cn.

design a lightweight network, which has only 0.89MB for a 3×352×352 input, to extract feature instead of VGGs or MobileNets. In order to retain the salient information of two modalities, we utilize a module named cross reference module (CRM) [17] to fuse the most salient information of depth features and RGB features. Subsequently, we utilize a feature refine module (FRM) to enhance the fused features. We use parallel dilated convolutions with different expansion rates to extract the large-scale information of fused features. Finally, considering the details of high-level features and the global semantic information of low-level features, a lightweight two-stage decoder is used to predict the salient object maps.

The main contributions of this paper can be summarized as follows:

- We propose an efficient two-stage decoder to combine different levels features. The decoder can fuse the detailed information of high-level features and the global semantic information of low-level features in two steps instead of top-down strategy.
- We design a feature refine module to enhance feature with larger receptive fields and channel attention. Four parallel dilated convolutions with different expansion rates can effectively extract the large-scale context information of features.
- We design a lightweight but efficient depth stream backbone instead of using the same backbone for RGB and depth. The customized backbone has fewer parameters but fits the model better.
- Compared with 10 state-of-the-art RGB-D SOD models on 5 datasets, our LTDNet shows outstanding performance both in terms of FPS and accuracy of evaluation indicators.

2. Related Work

2.1. Traditional RGB-D Salient Object Detection

The additional depth information is beneficial to more efficient and accurate localization and segmentation of salient objects in RGB images. Early methods utilize hand-crafted features for RGB-D SOD, such as boundary, contrast, shape attributes, 3D layout priors, anisotropic center-surround difference prior and so on. In [18], Peng et al. proposed a multi-contextual contrast model and built the first large-scale RGB-D dataset named NLPR for RGB-D SOD. In [19] Feng et al proposed local background enclosure features to solve the false positives due to areas of high contrast in background regions. In [20] Ren et al. obtained a saliency map by combining background, depth, region contrast, and orientation priors. In [21], Cong et al. proposed a depth-guided transformation model consisting of multilevel RGBD saliency initialization, depth-guided saliency refinement, and saliency optimization with depth constraints. However, traditional methods rely on hand-crafted features that lack high-level semantic representations and robustness in complex scenes.

2.2. Deep Learning-Based RGB-D Salient Object Detection

With the rapid development of deep learning, various methods based on convolutional neural networks (CNNs) have emerged. DF [22] is the first method to introduce deep learning into RGB-D SOD. Qu et al. fed hand-crafted features into a special-designed CNN model to fuse low-level salient features into hierarchical features and automatically detect salient objects in RGB-D images. In [23], Shigematsu et al. adopted two independent convolutional networks to process RGB images and hand-crafted deep features and fused the features to achieve salient maps. In CTMF [24], Han et al. utilized CNNs to transfer the structure of RGB images to be applicable for depth maps and fuses the high-level representations automatically to obtain saliency maps. In PCF [25], Chen et al. proposed a complementarity-aware fusion module to integrate complementary information from both modalities. In JL-DCF [10], taking depth map as a special case of RGB map, Fu et al. employed a shared CNN for RGB and depth feature extraction and presented joint learning and densely cooperative fusion to fuse multi-scale features effectively. In BBS-Net [13], Deng proposed a bifurcated backbone strategy to divide the multi-level features into teacher features and student features, and utilized a depth-enhanced module consisted of channel attention and spatial attention to enhance depth features. In S2MA [14] Liu et al. constructed a selective self-mutual attention module based on the non-local module to fuse multimodal information. In TANet [26], Chen et al. proposed a novel muti-modal fusion network from bottom-up and top-down perspectives, which introduces channel-wise attention mechanism to adaptively fuse complementary information from both modalities in each level. In D3Net [11], Fan et al. proposed a simple general three-stream RGB-D SOD architecture and designed a depth depurator unit which can filter low-quality depth maps.

Despite these methods have high accuracy, most of them possess large model size and computational cost.

2.3. Efficient RGB-D Salient Object Detection

In addition to the above-mentioned methods, some methods take computational cost and model efficiency into account. In A2dele [15], Piao et al. proposed a depth distiller constructed by network prediction and attention mechanisms to transfer depth information from depth stream to the RGB stream in training phrase, and only use RGB stream for SOD in testing phrase. In [27], Chen et al. designed a depth backbone which is much more efficient and lighter than traditional heavy backbones to learn feature representations. Although these models use special-designed measures, traditional backbones like VGGs and ResNets are still used to extract semantic information in RGB stream.

In recent years, more and more attention has been paid to the application of computer vision in mobile devices because of smart car, smart phone and intelligent robot. Due to the limited computational resources, traditional cumbersome backbones are not suitable for these mobile devices. A growing number of researchers focus on using efficient backbones like MobileNets and ShuffleNets as backbones for RGB-D SOD. In [28], Wu et al. used MobileNet-V2 as backbone to extract the feature of RGB images and a tailored network to extract the feature of depth maps. In [29], Huang et al. used ShuffleNet as feature extractors and middle-level feature fusion strategy to reduce the model size. In general, there is still little work in lightweight network for RGB-D SOD because of the weak feature extraction capability of the lightweight backbone

3. Proposed Method

3.1. Overview

The structure of the proposed LTDNet is shown in Fig 1, which mainly consists of four components: 1) RGB/Depth backbones; 2) cross reference module (CRM) 3) feature refine module (FRM). 4) two-stage decoder.

Following most RGB-D SOD methods, two individual branches s are respectively used to extract RGB image features and depth map features. The RGB branch is based on MobileNet-V2 [30] which discards the last maximum-pooling layer and fully connected layer, and the depth branch is a customized lightweight backbone which is based on inverted residual bottleneck blocks (IRB). Both branches output five features, and the output stride is 2 for each feature. We denote the output of RGB backbone and depth backbone as $\{R_1, R_2, R_3, R_4, R_5\}$ and $\{D_1, D_2, D_3, D_4, D_5\}$, respectively. The extracted features from different branches are sent to the CRM [17] to generate more informative features by mining and combining the most discriminative channels. We denote the fused features as $F_i\ (i = 1,2, \dots ,5)$. Then, the fused features are enhanced by extracting large-scale contextual semantic information through FRM, and the enhanced features are denoted as $E_i\ (i = 1,2, \dots ,5)$. Finally, in order to integrate the detailed information of high-level features and the global information of low-level features, the enhanced features are decoded in two steps to predict salient object.

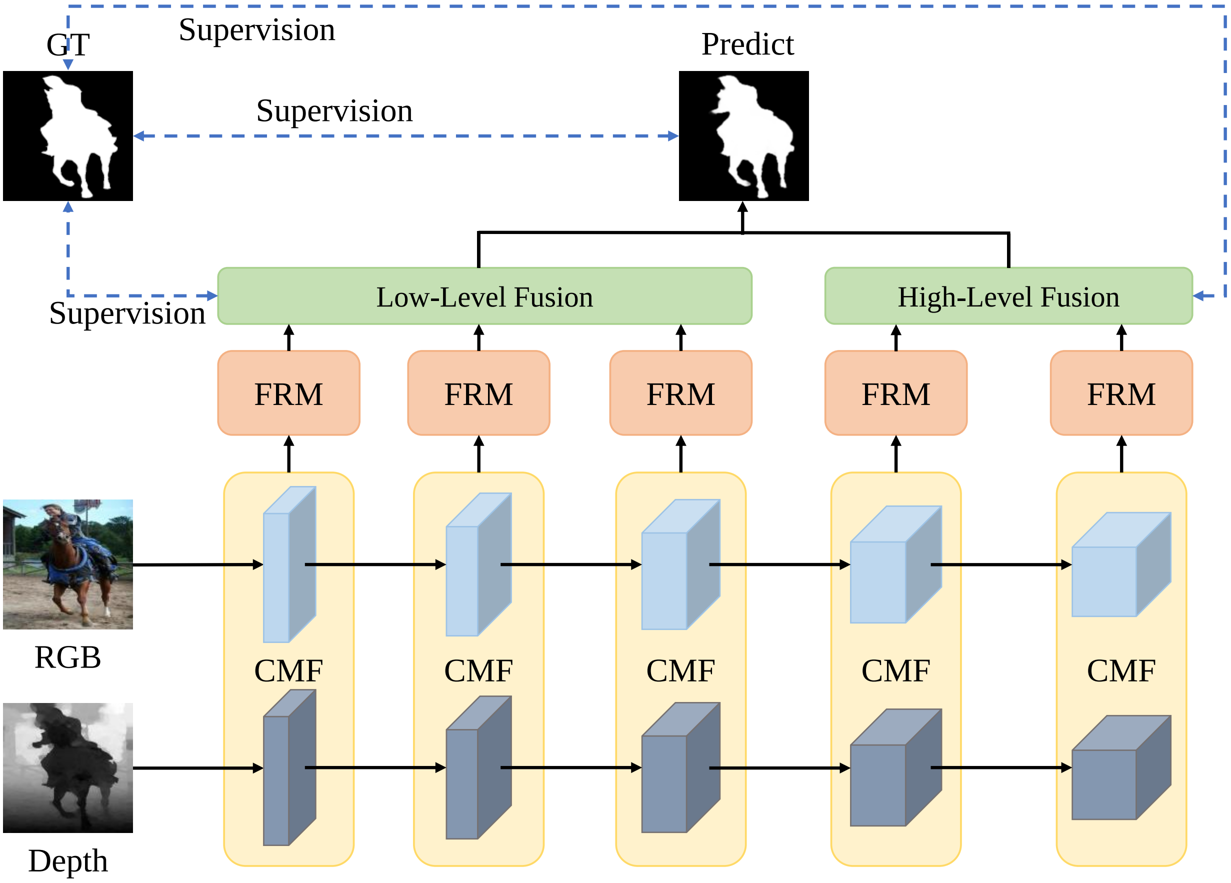

Fig 1. Architecture of the proposed LTDNet

3.2. Customized Depth Backbone

Table 1. Detail of the proposed customized depth backbone. About notations, t: expansion factor of IRB, c: output channels, n: number of block repeats, and s: stride of the first block.

Input	Output	Block	t	c	n	s
$352^2 \times 3$	$176^2 \times 16$	Conv2d	-	32	1	2
$176^2 \times 16$	$176^2 \times 16$	IRB	1	16	1	1
$176^2 \times 16$	$88^2 \times 24$	IRB	3	24	3	2
$88^2 \times 24$	$44^2 \times 32$	IRB	3	32	7	2
$44^2 \times 32$	$22^2 \times 96$	IRB	2	96	3	2
$22^2 \times 96$	$11^2 \times 320$	IRB	2	320	1	2

In general, depth have less information than RGB. Therefore, we design a lightweight network as the depth backbone instead of treating depth and RGB equally. The customized depth backbone is based on inverted residual bottleneck blocks (IRB) of MobileNet-V2, and the details is shown in Table 1.

Our customized depth backbone is much lighter than MobileNet-V2 (Ours: only 0.89Mb, MobileNet-V2: 13.82Mb), meanwhile it fits the proposed model better than MobileNet-V2 (see Sec. 4.5).

3.3. Cross Reference Module (CRM)

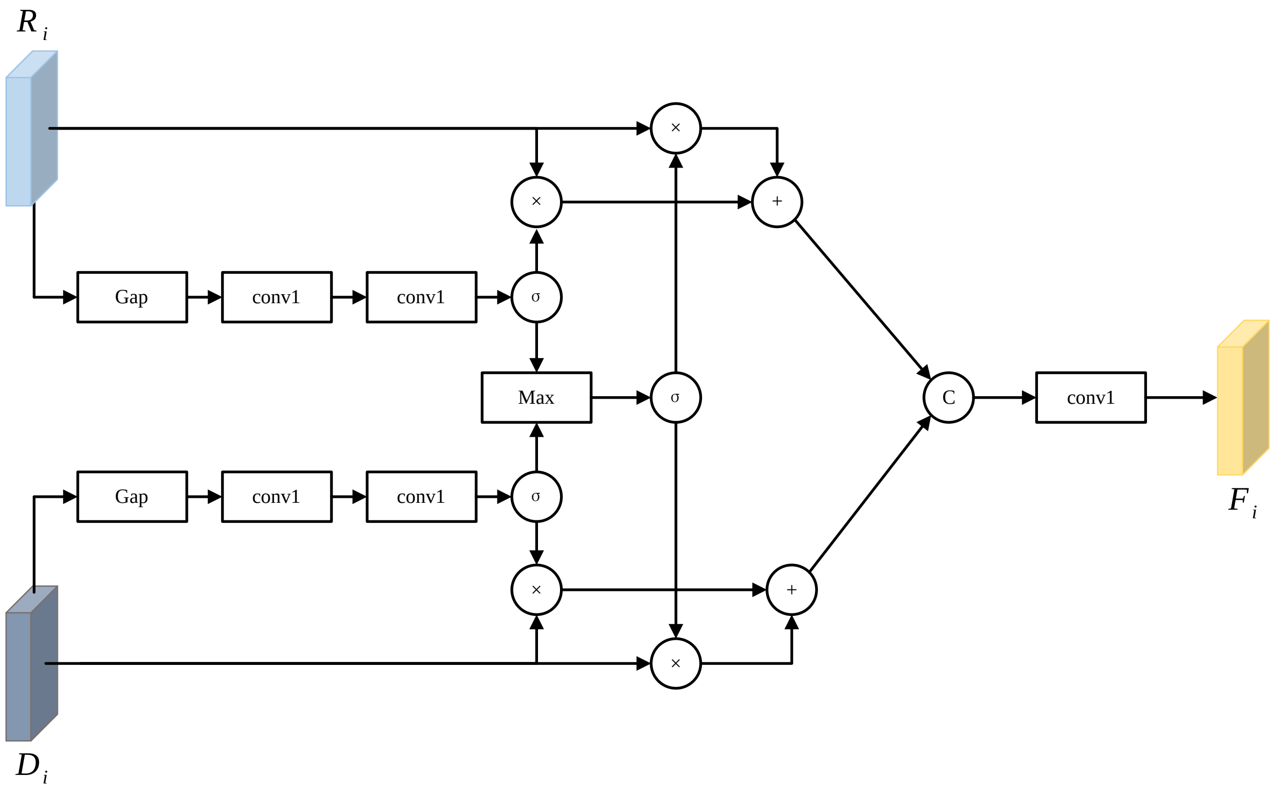

Fig 2. Architecture of CRM

In general, features from RGB contain rich texture and semantic information, and features from depth contain more discriminative scene layout clues. The features of two modalities are complementary each other. CRM [17] aims to generate more informative features by means of mining and combining cross-modality information, the detailed structure is shown in Fig 2.

Specifically, given two input features R_i and D_i from RGB backbone and depth backbone, the CRM utilizes a global average pooling to acquire the global information of two features respectively. Then, two fully connected layers and sigmoid activation function are followed to obtain the channel attention att_i^{rgb} and att_i^{depth}. The procedure can be defined as

$$att_i^{rgb} = \sigma(fc_2(fc_1(gap(R_i)))) \tag{3.1}$$

$$att_i^{depth} = \sigma(fc_2(fc_1(gap(D_i)))) \tag{3.2}$$

where $gap(\cdot)$ denotes global average pooling operation, $fc_1(\cdot)$ and $fc_2(\cdot)$ denote fully connected layers, $\sigma(\cdot)$ denotes sigmoid activation function, att_i^{rgb} and att_i^{depth} denote the channel attention of R_i and D_i respectively.

Then, the enhanced features of channel are obtained by channel-wise multiplication, formulated as

$$\bar{R}_i = att_i^{rgb} \otimes R_i \tag{3.3}$$

$$\bar{D}_i = att_i^{depth} \otimes D_i \tag{3.4}$$

where $\otimes$ denotes channel-wise multiplication operation.

The att_i^{rgb} and att_i^{depth} are aggregated by taking the maximum value of the corresponding channel to combine the most discriminative channel clues of R_i and D_i. Then, a sigmoid activation function is followed. The specific operation can be expressed as

$$att_i^{max} = \sigma(\max(att_i^{rgb}, att_i^{depth})) \tag{3.5}$$

where $Max(\cdot)$ indicates the maximum operation, and att_i^{max} is the cross-referenced channel attention.

Based on att_i^{max}, $\bar{R}_i$, $\bar{D}_i$, R_i and D_i, the features with the most discriminative information can be obtained by the following formulas:

$$\tilde{R}_i = \bar{R}_i + att_i^{max} \otimes R_i \tag{3.6}$$

$$\tilde{D}_i = \bar{D}_i + att_i^{max} \otimes D_i \tag{3.7}$$

where $\tilde{R}_i$ and $\tilde{D}_i$ are the enhanced features from different branches.

Finally, $\tilde{R}_i$ and $\tilde{D}_i$ are concatenated and fed into a 1×1 convolutional layer to fuse features from both modalities, the procedure can be denoted as

$$F_i = conv(cat(\tilde{R}_i, \tilde{D}_i)) \tag{3.8}$$

where $cat(\cdot)$ denotes channel concatenation, and F_i is the fused feature.

3.4. Feature Refine Module (FRM)

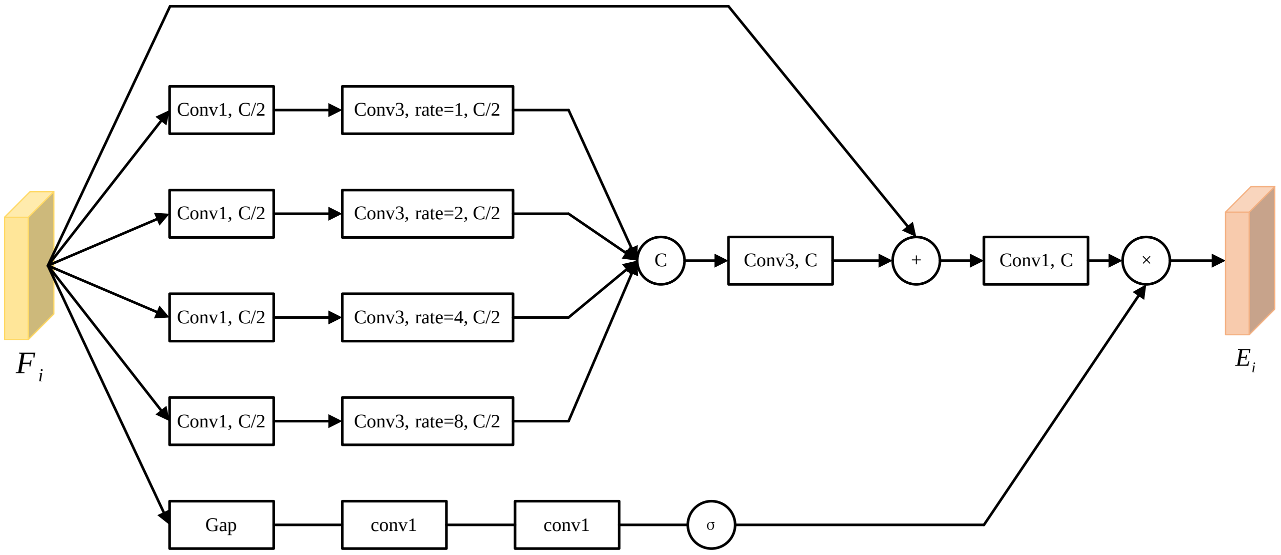

Fig 3. Architecture of FRM

Since the multi-stage pooling in backbones change the size and spatial structure of features, we design a feature refining module (FRM) to obtain multi-scale context information, which contains four parallel dilated convolutions with different expansion rates.

To improve efficiency and reduce parameters, FRM uses four parallel 1×1 convolutions to compress the channel of F_i to half. Then, four convolutions with dilation rates of 1, 2, 4, 8 are follows respectively for multi-scale information fusion. This procedure can be formulated as

$$F_i^{d1} = Conv_{3\times3}^{d1}(Conv_{1\times1}(F_i)) \tag{3.9}$$

$$F_i^{d2} = Conv_{3\times3}^{d2}(Conv_{1\times1}(F_i)) \tag{3.10}$$

$$F_i^{d3} = Conv_{3\times3}^{d3}(Conv_{1\times1}(F_i)) \tag{3.11}$$

$$F_i^{d4} = Conv_{3\times3}^{d4}(Conv_{1\times1}(F_i)) \tag{3.12}$$

where $d1$, $d2$, $d3$ and $d4$ are dilation rates of 1, 2, 4 and 8, $Conv_{1\times1}$ denotes 1×1 convolution, $Conv_{3\times3}^{di}$ $(i = 1,2,3,4)$ denotes 3×3 convolution with dilation rate of di and padding of di, F_i^{d1}, F_i^{d2}, F_i^{d3} and F_i^{d4} are the results obtained. Subsequently, F_i^{d1}, F_i^{d2}, F_i^{d3} and F_i^{d4} are concatenated along the channel axis and fed into a 3×3 convolution to fuse multi-scale features and compress the channel to the same as F_i. A residual connection is used and a 1×1 convolution followed for better optimization. The specific operations are formulated as

$$\bar{F_i} = Conv_{1\times1}(F_i + Conv_{3\times3}(cat(F_i^{d1}, F_i^{d2}, F_i^{d3}, F_i^{d4}))) \qquad (3.13)$$

where $Conv_{3\times3}$ denotes 3×3 convolution with padding of 1, $cat(\cdot)$ denotes concatenate operation. In order to uses global contextual information to realign the fused features $\bar{F_i}$, the attention mechanism is applied to F_i, the procedure can be defined as

$$E_i = att(F_i) \otimes \bar{F_i} \qquad (3.14)$$

where $att(\cdot)$ denotes the same operations as Equ. (1).

3.5. Two-stage Decoder

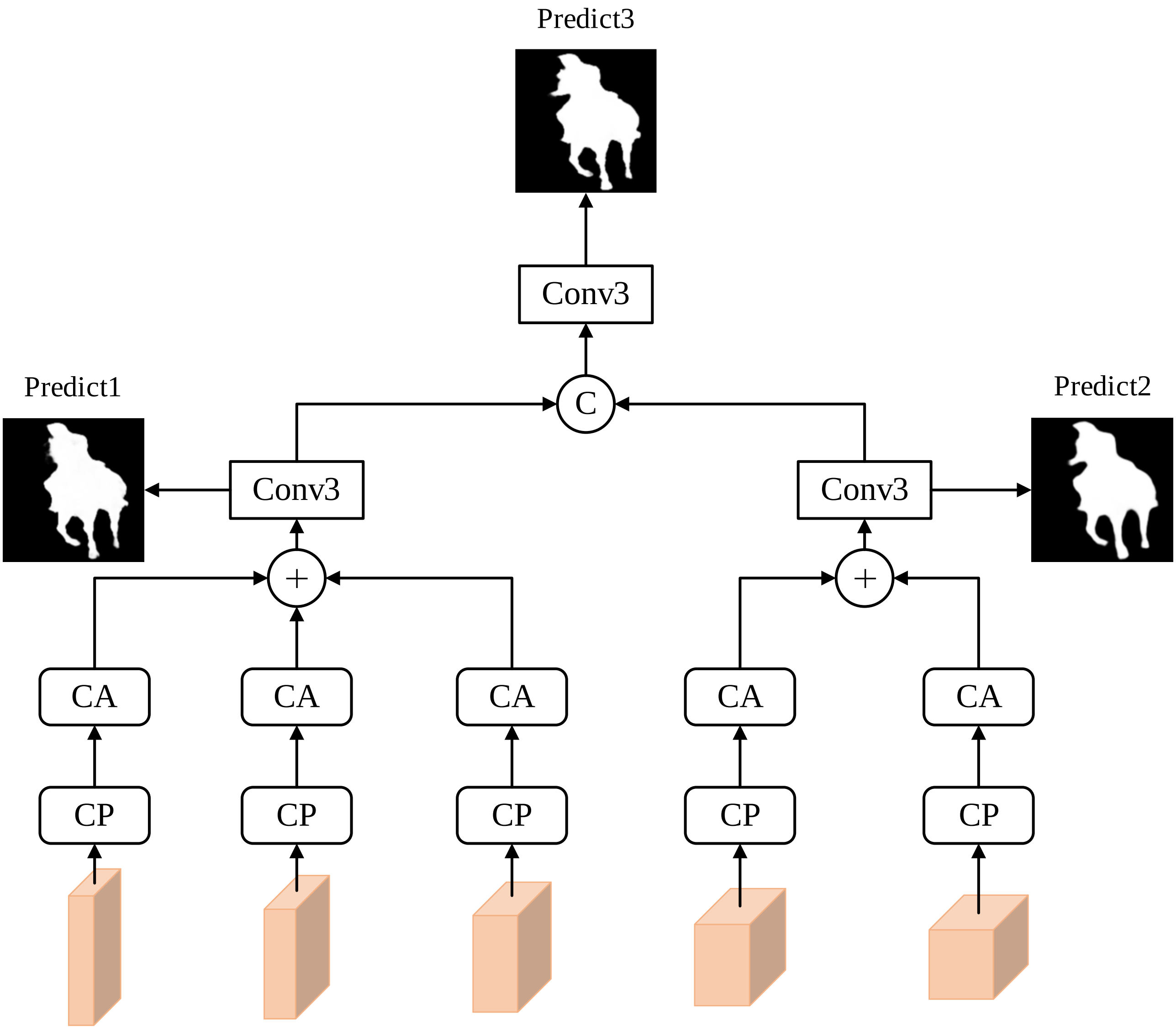

Fig 4. Architecture of Two-stage Decoder

Unlike most RGB-D SOD methods which apply top-down decoding strategy, we design an efficient decoder to integrate both global contextual information and detailed information in two steps. The first step is mainly to aggregate high-level features and low-level features separately to form two features. One has more details and the other focuses on global information. The details are shown in Fig 4.

In order to improve the efficiency of the whole decoding process, we utilize 3×3 convolutional (followed by BatchNorm layers and ReLU activation) to compress the channel of features passed from the FRM, and the compressed channels are 16, 16, 16, 32 and 32. Then, channel attention is utilized to enhance features. The process can be formulated as

$$\bar{E}_i = CA(CP(E_i)) \text{ (i=1,2,3,4,5)} \tag{3.15}$$

where $CP(\cdot)$ represents compress operation, $CA(\cdot)$ denotes channel attention as Equ. (1). Subsequently, we fuse the high-level features with the high-level features and the low-level features with the low-level features by element-wise addition, which produces two features containing different clues, formulated as

$$\bar{E}_{low} = \bar{E}_1 + up_{\times 2}(E_2) + up_{\times 4}(\bar{E}_3) \tag{3.16}$$

$$\bar{E}_{high} = \bar{E}_4 + up_{\times 2}(\bar{E}_5) \tag{3.17}$$

where $up_{\times 2}$ indicates bilinear upsampling by a factor of 2, and $up_{\times 4}$ indicates bilinear upsampling by a factor of 4. In the next step, we concatenate $\bar{E}_{low}$ and $\bar{E}_{high}$ and feed them into two 3×3 convolutions (with BatchNorm layers and ReLU activation) to promote the fusion, the procedure can be denoted as

$$E = Conv_{3\times 3}(Conv_{3\times 3}(cat(\bar{E}_{low}, up_{\times 8}(\bar{E}_{high})))) \tag{3.18}$$

where E denotes the fused feature of decoder. Finally, we feed the features obtained in the two stages into 3×3 convolutions whose out channel is 1 followed by a sigmoid activation respectively, and resize the saliency maps to the input images size, the procedure can be denoted as

$$P_1 = up_{\times 2}(Conv_{3\times 3}(\bar{E}_{low})) \tag{3.19}$$

$$P_2 = up_{\times 16}(Conv_{3\times 3}(\bar{E}_{high})) \tag{3.20}$$

$$P_3 = up_{\times 2}(Conv_{3\times 3}(E)) \tag{3.21}$$

where P_1, P_2 and P_3 is the final saliency maps.

3.6. Hybrid Loss Function

We apply a hybrid loss to optimize the network which consists of binary cross-entropy loss (BCE) [31], intersection over union loss (IoU) [32] and structural similarity index loss (SSIM) [33], formulated as

$$L_{P_i} = L_{BCE}(P_i, GT) + L_{IoU}(P_i, GT) + L_{SSIM}(P_i, GT) \tag{3.22}$$

where P_i represents the saliency map generated by decoder, GT denotes the ground truth, $L_{BCE}()$ denotes binary cross-entropy loss function, $L_{IoU}()$ denotes intersection over union loss function, $L_{SSIM}()$ denotes structural similarity index loss function, L_{P_i} denotes the loss of saliency map P_i. Then, the total loss can be expressed as

$$L_{total} = \sum_{i=1}^{3} \lambda_i L_{P_i} \tag{3.23}$$

where L_{total} represents the total loss, λ_i is a balance weight of L_{P_i}. Empirically, we set $\lambda_1 = 3$, $\lambda_2 = 2$, $\lambda_3 = 5$ to accelerate the convergence of loss.

4. Experiments

4.1. Datasets

We conducted our experiments on five public RGB-D benchmark datasets: NJU2K [34], NLPR [18], STERE [35], DES [36] and SIP [11]. NJU2K [34] contains 1985 groups of images collected from 3D movies, the Internet, and taken by Fuji W3 stereo cameras. NLPR [18] contains 1000 groups of images taken by Microsoft Kinect. STERE [35] contains 1000 groups of binocular images which consist of indoor and outdoor scenes. DES [36] consists of 135 groups of image captured by Microsoft Kinect in 7 indoor scenes and most of the scenes have a single salient object. SIP [11] consists of 929 high-resolution images captured by Huawei Mate10 with high quality depth maps.

Following [14,28,37-39], we use 700 images of NLPR [18] and 1500 images of NJU2K [34] for training, and the other 300 images of NLPR [18] and 485 images of NJU2K [34] for testing. Other datasets are used for testing.

4.2. Metrics

Following the works [13-14], we employ five commonly used metrics to evaluate the performance: mean absolute error (MAE) [40], maximum F-measure (F_{β}^{max}) [41], S-measure (S_{α}) [42], maximum E-measure (E_{ξ}^{max}) [43].

MAE [40] calculates the average absolute error between predicted saliency map S and the ground truth map GT, and it mainly evaluates the approximation between the predicted saliency map and the ground truth map. The specific calculation formula is

$$MAE = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} \left| S_{i,j} - GT_{i,j} \right| \tag{3.24}$$

where W and H represent the width and height of the image, respectively.

F-measure [41] is the harmonic mean of recall and precision, which is essentially a similarity measure based on regions. The specific calculation formula is:

$$F_{\beta} = \frac{(1+\beta^2) \times P \times R}{\beta^2 \times P + \mathrm{R}} \tag{3.25}$$

where P represents the precision, R represents the recall, and β represents the relative importance of recall to precision. We set $\beta^2 = 0.3$ as suggested in recent works.

S-measure [42] is a structural measure index which evaluates the structural similarity between the predicted saliency map and ground truth from regional perception and object perception. The specific calculation formula is as follows:

$$S_{\alpha} = \alpha S_o + (1 - \alpha) S_r \tag{3.26}$$

where S_o denotes regional perception, S_r denotes object perception, and $\alpha \in [0,1]$ is the balance parameter. We refer to [42] which set $\alpha = 0.5$ as default.

E-measure [43] is an enhanced-alignment measure that uses local-pixel values and image-level averages to obtain image statistics and local-pixel matching information. The specific calculation formula is as follows:

$$\xi_{FM} = \frac{2\varphi_{GT} \circ \varphi_{FM}}{\varphi_{GT} \circ \varphi_{GT} + \varphi_{FM} \circ \varphi_{FM}} \tag{3.27}$$

$$E_{\xi} = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} f(\xi_{FM}) \tag{3.28}$$

where $\circ$ donates the Hadamard product, φ_{GT} and φ_{FM} denotes bias matrices of ground-truth map GT and binary foreground map FM, $f(\cdot)$ represents a quadratic function.

4.3. Implementation Details

We implement experiments on a workstation with AMD Ryzen 5 5600X CPU, NVIDIA RTX 3060ti GPU with CUDA 11.0. PyTorch [44] toolbox is utilized to accelerated computing.

In order to prevent overfitting, data augmentation is performed including random horizontal flipping, random region cropping, random rotation, color enhancement and so on. After data augmentation, the image is resized to 352×352. We simply duplicate depth map into three channels as input to depth backbone.

During training, pre-trained MobileNet-v2 on ImageNet is used to initialize RGB backbone. We use Adam optimization to train LTDNet for 160 epochs. The initial learning rate is set as 1e-4, weight decay is 0.0005 and batch size is 10. Step learning rate policy is used to adjust learning rate that the learning rate is multiplied by 0.95 every 10 epochs.

4.4. Quantitative Comparison and Qualitative Comparison

We compare our method with 10 state-of-the-art (SOTA) RGB-D SOD methods, including PCF [25], MMCI [46], CPFP [38], DMRA [47], D3Net [11], SSF [48], FCMNet [49], DANet [39], ATST [50], A2dele[15]. For fair comparison, the codes and saliency maps are provided by the corresponding authors.

The results of the quantitative evaluation are shown in Table 2. We can clearly find that the proposed model has the smallest size while generally outperforming other methods in S-measure, F-measure, E-measure and MAE on the five datasets. Compared to A2dele who is the second smallest in these models, our model outperformed it on all datasets. Especially on the SIP dataset, S-measure, F-measure and E-measure increase by 5.9%, 6.6%, and 3.6% respectively, while MAE decreases by 27.1%. In addition, our model size is only about 1/3 of ATST whose performance is great, but on the NLPR dataset S-measure, F-measure, and E-measure increase by 2.2%, 4.8%, and 2.0% respectively. Meanwhile, MAE decrease by 14.3%. These quantitative comparisons demonstrate the superiority of LTDNet.

As shown in Fig 5, to further demonstrate the superiority of our model, we selected five models (8 groups of images for each method) in the above methods for visual comparison. Comparing the results in rows 3 and 4, we can see that our model is better at capturing salient object regions in scenes with multiple objects. Observing the results in rows 5 and 7, our model performs better when foreground and background are similar. In the rows 2 and 8, we know that our model has better salient object detection ability in c confused background. In general, LTDNet outperforms other methods in some complex scenes: multi-object scenes, similar foreground and background, confused background scenes and so on.

Table 2. Quantitative results compared with ten RGB-D SOD methods on five datasets. ↑/↓ indicates the larger/smaller, the better. The best results are highlighted in bold and red.

Metric		PCF	MMCI	CPFP	DMRA	D3Net	SSF	FCM	DANet	ATST	A2dele	Ours
		CVPR	PR	CVPR	ICCV	TNNI	CVPR	Scien	ECCV	ECCV	CVPR	-
		18	19	19	19	20	20	22	20	20	20	-
		[25]	[46]	[38]	[47]	[11]	[48]	[49]	[39]	[50]	[15]	-
Size(Mb)		534	930	278	228	530	125	197	102	123	57	**30**
SIP	S_α ↑	.842	.833	.850	.806	.860	.874	.858	**.878**	.864	.829	**.878**
	F_β^{max} ↑	.838	.818	.851	.821	.861	.880	.881	.884	.873	.834	**.889**
	E_ξ^{max} ↑	.901	.897	.903	.875	.909	**.921**	.912	.920	.911	.889	**.921**
	MAE ↓	.071	.086	.064	.085	.063	.053	.062	.054	.058	.070	**.051**
NJU2K	S_α ↑	.877	.858	.879	.886	.900	.899	.901	.891	.901	.868	**.908**
	F_β^{max} ↑	.872	.852	.877	.886	.900	.896	.907	.880	.893	.872	**.912**
	E_ξ^{max} ↑	.924	.915	.926	.927	**.950**	.935	.929	.932	.921	.914	.946
	MAE ↓	.059	.079	.053	.051	.041	.043	.044	.048	.040	.052	**.039**
NLPR	S_α ↑	.874	.856	.888	.899	.912	.914	.916	.915	.907	.890	**.927**
	F_β^{max} ↑	.841	.815	.867	.879	.897	.896	.908	.903	.876	.875	**.918**
	E_ξ^{max} ↑	.925	.913	.932	.947	.953	.953	.949	.953	.945	.937	**.964**
	MAE ↓	.044	.059	.036	.031	.025	.026	**.024**	.029	.028	.031	**.024**
DES	S_α ↑	.842	.848	.872	.900	.898	.905	.905	.904	.907	.884	**.919**
	F_β^{max} ↑	.804	.822	.846	.888	.885	.883	.913	.894	.885	.873	**.914**
	E_ξ^{max} ↑	.893	.928	.923	.943	.946	.941	.949	.957	.952	.920	**.959**
	MAE ↓	.049	.065	.038	.030	.031	.025	.025	.029	.024	.030	**.022**
STERE	S_α ↑	.875	.873	.879	.835	.899	.893	.899	.892	.897	.885	**.903**
	F_β^{max} ↑	.860	.863	.874	.847	.891	.890	**.904**	.881	.884	.885	.900
	E_ξ^{max} ↑	.925	.927	.925	.911	.939	.936	.939	.930	.921	.935	**.944**
	MAE ↓	.064	.068	.051	.066	.046	.044	.043	.048	**.039**	.043	.040

4.5. Ablation Studies

In this section, ablation experiments are conducted, mainly studies: (1) the applicability of the customized depth backbone; (2) the necessity of CRM; (3) the impact of FRM on LTDNet. The above three problems are analyzed by ablation. We analyze the above problems in terms of four evaluation indexes and visual comparison.

To study the impact of the customized depth backbone, we replace the customized backbone with MobileNet-V2. It is not difficult to find that compared with "Ours", S-measure, max F-measure, and max E-measure decrease slightly on five datasets in "-Backbone", while the MAE increase. For example, the S-measure, max F-measure, and max E-measure decrease by 1.94%, 2.81%, and 2.06% respectively on the SIP dataset, while the MAE increases by 13.73%. This indicates that customized depth backbone s more suitable for LTDNet.

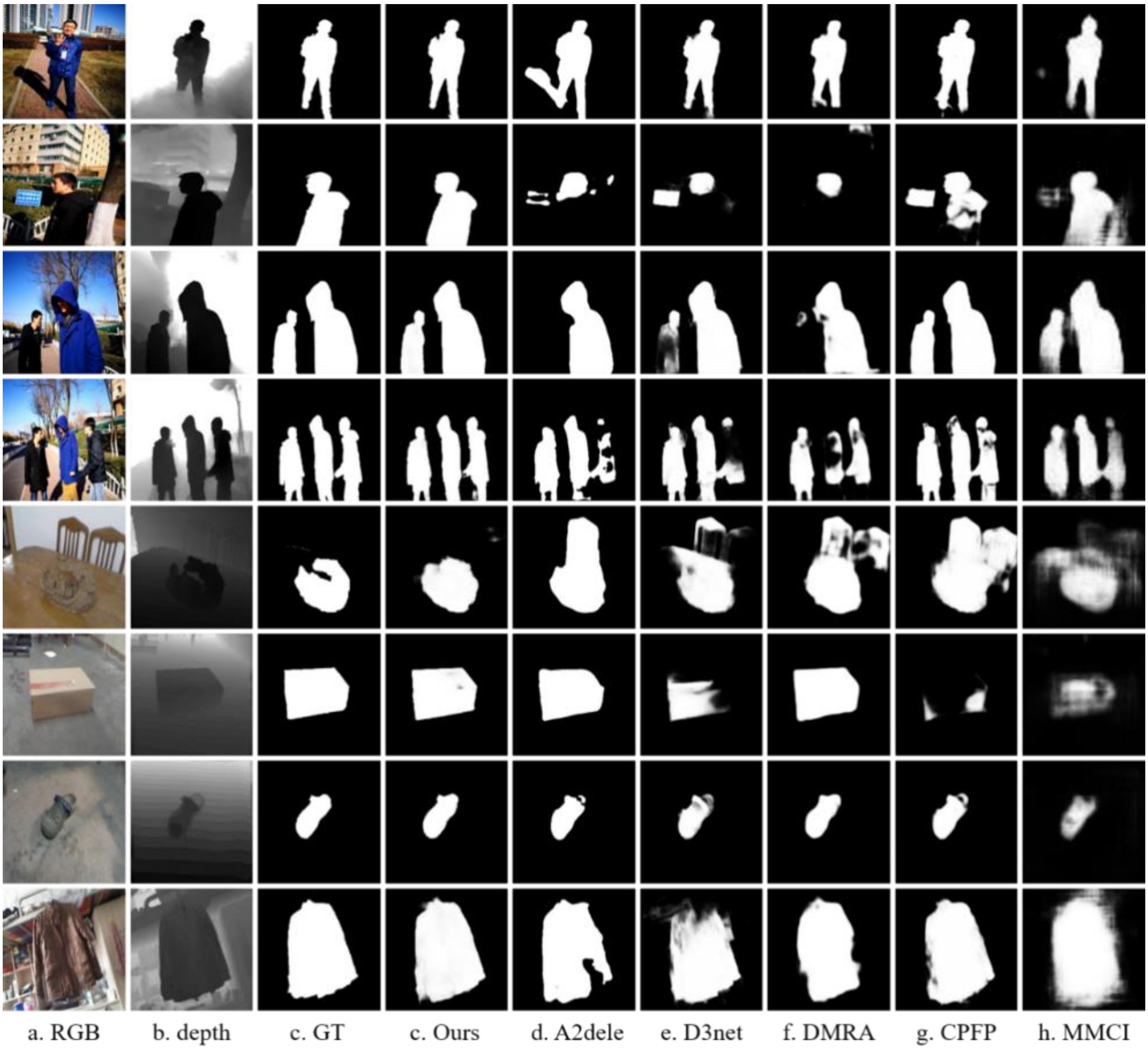

Fig 5. Visual comparisons with SOTA methods

Table 3. Ablation study of LTDNet. -backbone: uses MobileNet-V2 to replace the customized depth backbone. -CRM(C): use concatenation and convolution to replace CRM. -CRM(+): use element-wise addition and convolution to replace CRM. -FRM: remove FRM from the model. -FRM(RFB): use RFB to replace FRM.

Metric		-Backbone	-CRM(C)	-CRM(+)	-FRM	-FRM(RFB)	Ours
SIP	$S_\alpha \uparrow$	.861	.880	.875	.872	.869	.878
	$F_\beta^{max} \uparrow$	.864	.883	.881	.877	.869	.889
	$E_\xi^{max} \uparrow$	.902	.922	.917	.915	.913	.921
	$MAE \downarrow$	.058	.049	.051	.055	.055	.051
NJU2K	$S_\alpha \uparrow$	.900	.899	.898	.895	.903	.908
	$F_\beta^{max} \uparrow$	.900	.898	.898	.896	.900	.912
	$E_\xi^{max} \uparrow$	.937	.940	.937	.937	.942	.946
	$MAE \downarrow$	.042	.042	.041	.045	.041	.039
NLPR	$S_\alpha \uparrow$	.925	.821	.921	.912	.922	.927
	$F_\beta^{max} \uparrow$	.913	.910	.911	.895	.911	.918
	$E_\xi^{max} \uparrow$	.958	.958	.959	.949	.958	.964
	$MAE \downarrow$	.025	.025	.024	.030	.025	.024
DES	$S_\alpha \uparrow$	.915	.919	.899	.929	.916	.919
	$F_\beta^{max} \uparrow$	.907	.912	.876	.926	.907	.914
	$E_\xi^{max} \uparrow$	.953	.958	.928	.968	.953	.959
	$MAE \downarrow$	.023	.022	.027	.021	.024	.022
STERE	$S_\alpha \uparrow$	.897	.900	.900	.893	.899	.903
	$F_\beta^{max} \uparrow$	.890	.895	.894	.892	.892	.900
	$E_\xi^{max} \uparrow$	.934	.940	.938	.939	.939	.944
	$MAE \downarrow$	.044	.041	.041	.044	.042	.040

In the necessity analysis of CRM, we set up two groups of experiments: (1) replace CRM with element-wise addition and 3×3 convolution ("-CRM(+)" in Table 3), (2) replace CRM with concatenation operation and 3×3 convolution ("-CRM(C)" in Table 3). Looking at Table 3, we find that the performance of the two groups of ablation experiments is slightly worse than that of the model in this paper (except for the S-measure and E-measure of the SIP dataset in the "CRM(C)" group). For example, the S-measure, max F-measure, and max E-measure decrease by 2.18%, 4.16%, and 3.23% respectively on the DES dataset, while the MAE increases by 22.73%. This demonstrates that the use of CMR can generally enhance the performance of LTDNet.

In the ablation analysis of FRM, we also set up two groups of experiments: (1) remove FRM ("-FRM" in Table 3); (2) replace FRM module with RFB [51]. Comparing "-FRM" and "Ours", it is easy to find that the performance of the model is generally reduced after removing the FRM (except for the DES dataset). In addition, the performance is all degraded after

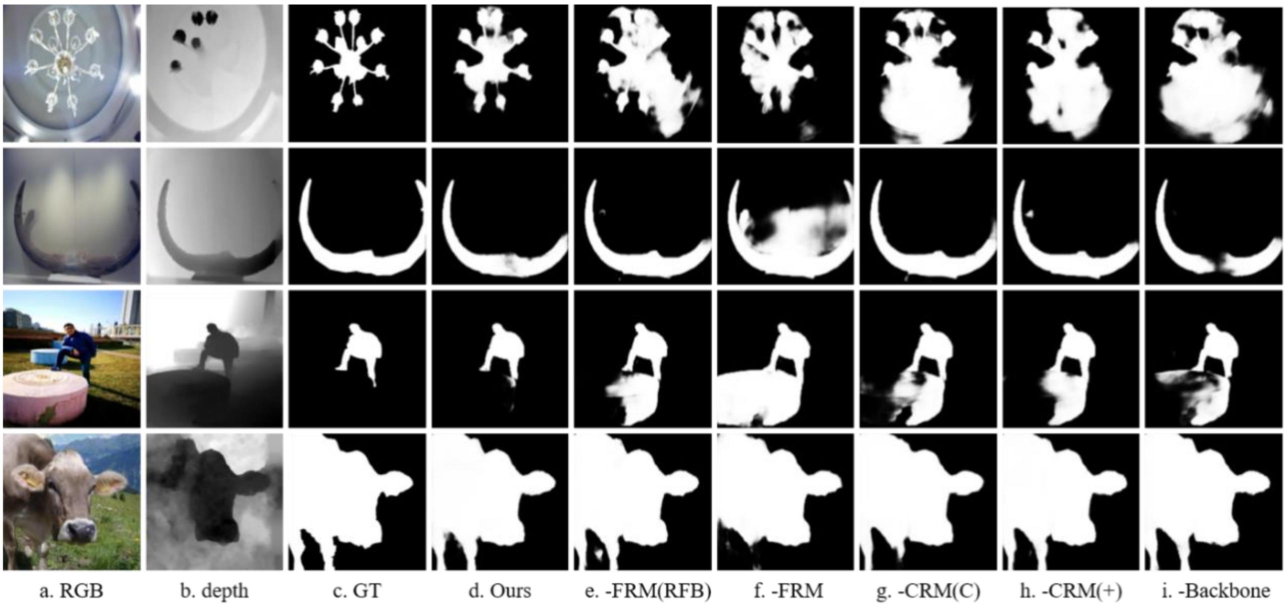

Fig 6. Visual comparisons of ablation experiments

replacing FRM with RFB module. For example, the S-measure, max F-measure, and max E-measure decrease by 1.03%, 2.25%, and 0.87% respectively on the SIP dataset, while the MAE increases by 7.84%. The experiments show that FRM can improve the performance of LTDNet in most senses, and is more suitable for our model than others modules like RFB.

As shown in Fig 6, we selected 4 groups of images from ablation experiments for comparison. We clearly find that each component in LTDNet is reasonable and efficient.

5. Conclusion

In this paper, we propose an efficient model called LTDNet. Considering the problem of expensive computational cost brought by lumbersome models, we use lightweight backbones to replace traditional cumbersome backbones. MobileNet-V2 is utilized to extract features from RGB maps and an efficient backbone is designed to extract features from depth maps. To mine the most discriminative information, we use CRM to fuse cross-modality clues to achieve complementarity between RGB and depth. In addition, we use FRM, which has a large receptive field, to enhance the fused features. Finally, a lightweight two-stage decoder is used to obtain saliency maps. Experiments on 5 benchmark datasets show that the proposed model has the fewest parameters, while performing better than SOTA methods.

References

1. Hong S, You T, Kwak S, et al. Online tracking by learning discriminative saliency map with convolutional neural network[C]//International conference on machine learning. PMLR, 2015: 597-606.
2. Tsai C C, Li W, Hsu K J, et al. Image co-saliency detection and co-segmentation via progressive joint optimization[J]. IEEE Transactions on Image Processing, 2018, 28(1): 56-71.
3. Adjabi I, Ouahabi A, Benzaoui A, et al. Past, present, and future of face recognition: A review[J]. Electronics, 2020, 9(8): 1188.
4. Fan D P, Ji G P, Zhou T, et al. Pranet: Parallel reverse attention network for polyp segmentation[C]//International conference on medical image computing and computer-assisted intervention. Springer, Cham, 2020: 263-273.
5. Piao Y, Ji W, Li J, et al. Depth-induced multi-scale recurrent attention network for saliency detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7254-7263.

6. Qin X, Zhang Z, Huang C, et al. Basnet: Boundary-aware salient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 7479-7489.

7. Li G, Liu Z, Ling H. ICNet: Information conversion network for RGB-D based salient object detection[J]. IEEE Transactions on Image Processing, 2020, 29: 4873-4884.

8. Zhang Q, Huang N, Yao L, et al. RGB-T salient object detection via fusing multi-level CNN features[J]. IEEE Transactions on Image Processing, 2019, 29: 3321-3335.

9. Tu Z, Xia T, Li C, et al. RGB-T image saliency detection via collaborative graph learning[J]. IEEE Transactions on Multimedia, 2019, 22(1): 160-173.

10. Fu K, Fan D P, Ji G P, et al. JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 3052-3062.

11. Fan D P, Lin Z, Zhang Z, et al. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks[J]. IEEE Transactions on neural networks and learning systems, 2020, 32(5): 2075-2089.

12. Li G, Liu Z, Ye L, et al. Cross-modal weighting network for RGB-D salient object detection[C]//European Conference on Computer Vision. Springer, Cham, 2020: 665-681.

13. Fan D P, Zhai Y, Borji A, et al. BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network[C]//European conference on computer vision. Springer, Cham, 2020: 275-292.

14. Liu N, Zhang N, Han J. Learning selective self-mutual attention for RGB-D saliency detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 13756-13765.

15. Piao Y, Rong Z, Zhang M, et al. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9060-9069.

16. Wang X, Li S, Chen C, et al. Data-level recombination and lightweight fusion scheme for RGB-D salient object detection[J]. IEEE Transactions on Image Processing, 2020, 30: 458-471.

17. Ji W, Li J, Yu S, et al. Calibrated RGB-D salient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2021: 9471-9481.

18. Peng H, Li B, Xiong W, et al. RGBD salient object detection: A benchmark and algorithms[C]//European conference on computer vision. Springer, Cham, 2014: 92-109.

19. Feng D, Barnes N, You S, et al. Local background enclosure for RGB-D salient object detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 2343-2350.

20. Ren J, Gong X, Yu L, et al. Exploiting global priors for RGB-D saliency detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition workshops. 2015: 25-32.

21. Cong R, Lei J, Fu H, et al. Going from RGB to RGBD saliency: A depth-guided transformation model[J]. IEEE transactions on cybernetics, 2019, 50(8): 3627-3639.

22. Qu L, He S, Zhang J, et al. RGBD salient object detection via deep fusion[J]. IEEE transactions on image processing, 2017, 26(5): 2274-2285.

23. Shigematsu R, Feng D, You S, et al. Learning RGB-D salient object detection using background enclosure, depth contrast, and top-down features[C]//Proceedings of the IEEE international conference on computer vision workshops. 2017: 2749-2757.

24. Han J, Chen H, Liu N, et al. CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion[J]. IEEE transactions on cybernetics, 2017, 48(11): 3171-3183.

25. Chen H, Li Y. Progressively complementarity-aware fusion network for RGB-D salient object

detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 3051-3060.

26. Chen H, Li Y. Three-stream attention-aware network for RGB-D salient object detection[J]. IEEE Transactions on Image Processing, 2019, 28(6): 2825-2835.
27. Chen S, Fu Y. Progressively guided alternate refinement network for RGB-D salient object detection[C]//European Conference on Computer Vision. Springer, Cham, 2020: 520-538.
28. Wu Y H, Liu Y, Xu J, et al. MobileSal: Extremely efficient RGB-D salient object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021.
29. Huang N, Zhang Q, Han J. Middle-level Fusion for Lightweight RGB-D Salient Object Detection[J]. arXiv preprint arXiv:2104.11543, 2021.
30. Sandler M, Howard A, Zhu M, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2018: 4510-4520.
31. De Boer P T, Kroese D P, Mannor S, et al. A tutorial on the cross-entropy method[J]. Annals of operations research, 2005, 134(1): 19-67.
32. Máttyus G, Luo W, Urtasun R. Deeproadmapper: Extracting road topology from aerial images[C]//Proceedings of the IEEE international conference on computer vision. 2017: 3438-3446.
33. Wang Z, Simoncelli E P, Bovik A C. Multiscale structural similarity for image quality assessment[C]//The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. Ieee, 2003, 2: 1398-1402.
34. Ju R, Ge L, Geng W, et al. Depth saliency based on anisotropic center-surround difference[C]//2014 IEEE international conference on image processing (ICIP). IEEE, 2014: 1115-1119.
35. Niu Y, Geng Y, Li X, et al. Leveraging stereopsis for saliency analysis[C]//2012 IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2012: 454-461.
36. Cheng Y, Fu H, Wei X, et al. Depth enhanced saliency detection method[C]//Proceedings of international conference on internet multimedia computing and service. 2014: 23-27.
37. Zhang J, Fan D P, Dai Y, et al. UC-Net: Uncertainty inspired RGB-D saliency detection via conditional variational autoencoders[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 8582-8591.
38. Zhao J X, Cao Y, Fan D P, et al. Contrast prior and fluid pyramid integration for RGBD salient object detection[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 3927-3936.
39. Zhao X, Zhang L, Pang Y, et al. A single stream network for robust and real-time RGB-D salient object detection[C]//European Conference on Computer Vision. Springer, Cham, 2020: 646-662.
40. Borji A, Cheng M M, Jiang H, et al. Salient object detection: A benchmark[J]. IEEE transactions on image processing, 2015, 24(12): 5706-5722.
41. Achanta R, Hemami S, Estrada F, et al. Frequency-tuned salient region detection[C]//2009 IEEE conference on computer vision and pattern recognition. IEEE, 2009: 1597-1604.
42. Fan D P, Cheng M M, Liu Y, et al. Structure-measure: A new way to evaluate foreground maps[C]//Proceedings of the IEEE international conference on computer vision. 2017: 4548-4557.
43. Fan D P, Gong C, Cao Y, et al. Enhanced-alignment measure for binary foreground map evaluation[J]. arXiv preprint arXiv:1805.10421, 2018.
44. Paszke A, Gross S, Massa F, et al. Pytorch: An imperative style, high-performance deep learning library[J]. Advances in neural information processing systems, 2019, 32.
45. Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J].

Communications of the ACM, 2017, 60(6): 84-90.

46. Chen H, Li Y, Su D. Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection[J]. Pattern Recognition, 2019, 86: 376-385.

47. Piao Y, Ji W, Li J, et al. Depth-induced multi-scale recurrent attention network for saliency detection[C]//Proceedings of the IEEE/CVF International Conference on Computer Vision. 2019: 7254-7263.

48. Zhang M, Ren W, Piao Y, et al. Select, supplement and focus for RGB-D saliency detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 3472-3481.

49. Jin X, Guo C, He Z, et al. FCMNet: Frequency-aware cross-modality attention networks for RGB-D salient object detection[J]. Neurocomputing, 2022, 491: 414-425.

50. Zhang M, Fei S X, Liu J, et al. Asymmetric two-stream architecture for accurate RGB-D saliency detection[C]//European Conference on Computer Vision. Springer, Cham, 2020: 374-390.

51. Liu S, Huang D. Receptive field block net for accurate and fast object detection[C]//Proceedings of the European conference on computer vision (ECCV). 2018: 385-400.