Neurocomputing 454 (2021) 168–177

Contents lists available at ScienceDirect

Neurocomputing

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

A cross-modal edge-guided salient object detection for RGB-D image

Zhengyi Liu *, Kaixun Wang, Hao Dong, Yuan Wang

Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei, China

ARTICLE INFO

Article history:
Received 26 June 2020
Revised 7 February 2021
Accepted 5 May 2021
Available online 11 May 2021
Communicated by Zidong Wang

Keywords:
Salient object detection
Edge guidance
Cross modal
Gated fusion

ABSTRACT

Salient object detection simulates the attention mechanism of human behavior to grasp the most attractive objects in the images. Recently edge information has been introduced to enhance the sharp contour in RGB image saliency detection. Inspired by it, we probe into the edge-guided RGB-D image saliency detection. There are two key problems need to be solved. One is how to extract edge information from cross-modal color and depth information, the other is how to fuse the edge feature into double-stream saliency detection network. To solve these two issues, a cross-modal edge-guided salient object detection for RGB-D image is proposed. Based on double-stream U-Net framework, edge information is extracted from the deep and shallow block of both modalities. The feature in deep layer contains sematic information implying where are the object boundaries, so the features of both modalities are directly fused. The feature in shallow layer provides more detailed spatial information, so a gated fusion layer is utilized to fuse the features of both modalities to filter out the depth image noise. Extracted edge feature is fed into decoder combining with color and depth feature to achieve edge-guided cross-modal decoding process. Experimental results show our model outperforms SOTA models based on the edge guidance and gated fusion strategies in cross-modal double-stream network.

1. Introduction

Salient object detection simulates the attention mechanism of human behavior to grasp the most attractive objects in the images. It has been widely used in computer vision and robot systems, such as image segmentation [1], object recognition [2], visual question answering [3], visual tracking [4], and so on. It benefits from the powerful feature extraction capabilities of convolutional neural networks [5,6] to broken the limits of early research depending on prior knowledge. But the continuous convolution and pooling operations make the features in the deeper layer locate salient regions more accurately, but reduce the size of feature maps, which is not suitable for pixel-level salient object detection task. Although U-Net [7] and FPN[8] can progressively restore the size of feature maps by upsampling the feature in the higher layer and short connection between encoder and decoder, the boundary of the salient object is still indistinct. Therefore, sharpening the boundary of the salient objects by the edge detection appears. Some methods [9–11] use loss function to penalize errors on the boundary, some methods [12–14] introduce the external edge detection dataset to train edge feature, some methods [15,11, 16–22] joinly train salient object detection and edge detection to promote both performances, some methods [23–25] fuse edge information to enhance salient object detection.

Although combined with edge information, salient object detection still suffers the trouble of predicting in the cluttered background. Fortunately depth information is a useful supplement to the color information. As shown in Fig. 1, the saliency maps generated by PoolNet [12] without depth information is interior to ours with depth information in object integrity.

Therefore how to detect the salient objects by combining edge information in RGB-D images is our purpose. There are two key problems need to be solved compared with edge-guided salient object detection task for RGB image. One is how to extract the edge information from RGB-D images, the other is how to fuse edge information into salient object detection. SSF [22] explores edge details from high-levels of the RGB modal, and fuses edge information with color and depth information, respectively. cmSalGAN [25] introduces three residual convolutional blocks into the first three convolutional blocks of the encoder in RGB stream to extract edge information, and fuses them into the fusion result of RGB and depth streams. The edge information is only generated from color modality, and the performance improvement in experiments is not obvious. So we probe into better combining color and depth cross-modal information to extract edge information and further integrating edge information into cross-modal double streams. To be specific, since more detailed edge information is retained in the

* Corresponding author.
E-mail address: liuzywen@ahu.edu.cn (Z. Liu).

https://doi.org/10.1016/j.neucom.2021.05.013
0925-2312/

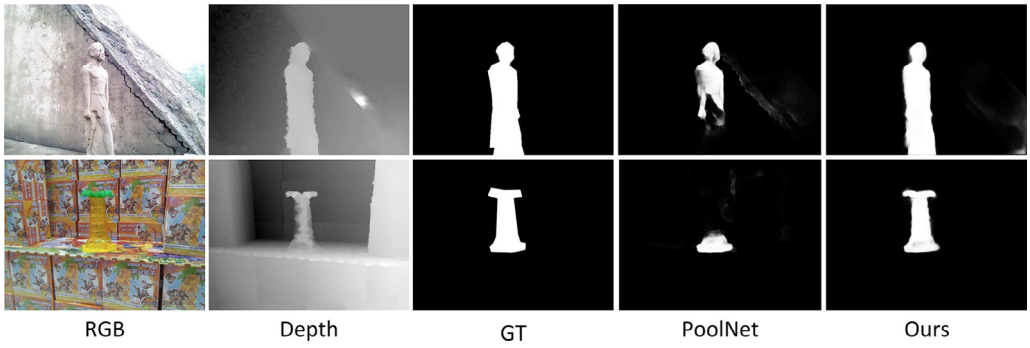


Fig. 1. Illustration of the importance of depth information for saliency detection.

feature of the shallow layer and more accurate object position information is reflected by the feature in the deep layer, edge information is extracted from the combination of the feature in the shallow layer and deep layer. Further, the feature in the shallow layer is noisy, so a gated fusion layer is adopted in the shallow layer to adaptively filter depth noise. After generating the edge information, we feed it back to guide the salient features to obtain the clearer object contour. As shown in Fig. 2, edge information is important in improving the performance of salient object detection. The saliency map is blurred when not using edge information (5^{th} column), while it becomes sharp when embedding the edge information (6^{th} column).

Our main contributions can be summarized as follows:

- Cross-modal edge-guided salient object detection model for RGB-D images is proposed. It extracts edge information from cross-modal color and depth information, and further fuses edge information into cross-modal color and depth features to generate the saliency map with clear boundary.
- When extracting cross-modal edge information, the cross-level color and depth features from the first block and the fifth block are utilized. Cross-modal color and depth features in the fifth block are directly fused, while those in the first block are fused by a gated fusion layer, to avoid the influence from the noise of depth feature in the shallow layer. Experimental result demonstrates it can filter out the depth noise.
- When fusing edge information into cross-modal decoding process, edge feature is first supervised by the salient edge ground truth, and then combined with cross-modal fused color and depth feature, and last to generate the saliency map which is supervised by the saliency map ground truth. Experimental result demonstrates edge-guided decoding process significantly improves the performance.
- Compared with the 9 STOA methods on seven widely used public benchmark datasets, our method has achieved good performance in quantitative and qualitative evaluation.

2. Related work

2.1. Salient object detection

Salient object detection mimics human attention mechanism and detects the most attractive objects. It has been developed from RGB saliency detection [26,24,27–32] to RGB-D saliency detection [33–37], co-saliency detection[38–43], video saliency detection [44–47], light-field SOD [48,49], high-resolution image SOD [50,51] and so on. Recently more and more modern smart phones such as iPhone X, Huawei Mate10, and Samsung Galaxy S10 have provided depth information which can improve the accuracy of salient object detection. The potential commercial application has given rise to the research tide of RGB-D image object-aware saliency detection.

Early researches [52–57] mainly considered depth contrast, and generated saliency maps by depth difference between regions. Different priors were then proposed to extract depth information, for example center-surround prior [58], boundary prior [59], surface orientation [60], background prior [61] and center-dark channel prior [62]. But the prior knowledge constructs the hand-crafted low-level features, it will not be always effective in different cluttered scenes. Fusion with mutual manifold ranking [63], saliency evolution strategy [64], location and two stage boundary refinement framework [56] and regressor [65] were adopted to overcome the weak of feature representation.

With the development of Convolutional Neural Networks (CNN) [5] and U-Net framework [7], the per-pixel end-to-end deep architecture has been used to detect the salient object in RGB-D images.

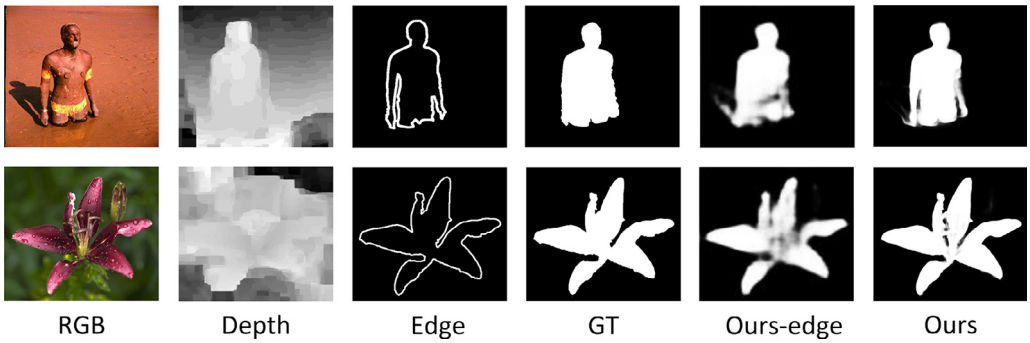


Fig. 2. Illustration of the importance of edge information for saliency detection.

With the increase of the network receptive field, the position of salient objects becomes more and more accurate. However, at the same time, spatial coherence is ignored. Although short-connection between encoder and decoder can progressively restore the size of feature map, the boundary of salient object is still indistinct. Jiang et al. [25]and Zhang et al. [22] introduced edge prior information into RGB-D saliency detection to solve the problem of boundary dilution, and extracted edge feature from RGB modality. We probe into extracting edge information from cross-modal RGB-D image and combining them to improve the RGB-D saliency detection.

2.2. Salient edge detection

In order to retain the structure information which is important for SOD, the early researches use the superpixel [66] for pre-processing or CRF [67–69] for post-processing. They need more time cost to preserve the object boundaries. Recently more and more end-to-end networks are proposed to detect the salient objects based on the edge information. Luo et al. [9] used an extra IoU-based edge loss to directly optimize the edges of predicted saliency maps. Chen et al. [10] used contour loss to perceive salient object boundaries. Wang et al. [11] used focal loss to facilitate the learning of the hard boundary pixels. Liu et al. [12] jointly and alternatively trained the edge detection task and salient object detection task using extra edge detection dataset. Wu et al. [13] trained saliency detection networks by exploiting the supervision from salient object detection, foreground contour detection and edge detection. Guan et al. [14] trained an edge detection network based on HED framework [70] to extract the boundary information, and integrated the edge contours with the saliency detection decoder to depict continuous boundary for salient objects. Tu et al. [71] designed an edge guidance block to embed edge prior knowledge [72] into hierarchical feature maps for effective feature representations. Zhang et al. [20] jointly learned salient edges and saliency labels in an end-to-end fashion. Extra hand-craft edge features were used as a complementary to preserve edge information effectively. Zhuge et al. [15] trained two networks for salient object detection and edge detection, and fused them by an attention-based feature fusion module. Wang et al. [11] trained two networks for salient object detection and edge detection and connected the decoders of both boundary and mask sub-networks. Wu et al. [16] trained a shared encoder and two parallel decoders for salient object detection and edge detection, and fused them in bidirectional integration manner. Zhou et al. [17] designed a multi-stage siamese network to parallelled estimate the salient maps of edges and regions at the same time, and the salient edge and salient region were concatenated as the edge-guided saliency map, which was used to predict the saliency map in the next stage. Lin et al. [19] simultaneously modelled the complementary information of saliency and boundary in a single network. By recurrently stacking strategy, the saliency features and boundary features will be progressively refined at the same time. Su et al. [18] designed transition compensation besides boundary localization and interior perception. It amended the probable failures between boundaries and interiors in a boundary-aware feature mosaic selection manner. Wang et al. [21] emphasized on the detection of salient edge information besides salient object detection, which can be leveraged for sharpening the salient object. Amulet [23] incorporated edge-aware feature maps in low-level layers and the predicted results from low resolution features to achieve accurate object boundary inference and semantic enhancement. EGNet [24] modelled and fused the complementary salient edge information and salient object information within a single network in an end-to-end manner.

Edge-guided salient object detection has been deeply explored in RGB images. In RGB-D salient object detection, Jiang et al. [25] and Zhang et al. [22] extracted edge prior information from RGB stream. Depth information is not well developed and utilized in edge extraction. We attempt the cross-modal edge extraction and further edge-guided salient object detection.

3. Proposed method

In this section, we will introduce our proposed cross-modal edge-guided network, which is shown in Fig. 3. It consists of double-stream encoder with attention mechanism, cross-modal and cross-level salient edge detector and edge-guided cross-modal decoder. Our model adopts double-stream U-Net structure based on ResNet-50 with CBAM. Cross-modal and cross-level salient edge detector is proposed to extract the better edge information from color and depth modalities. Edge-guided cross-modal decoder adds edge guidance besides cross-modal progressive upsampling fusion process in the decoding process.

3.1. Double-stream encoder with attention mechanism

RGB image shows the color distribution, texture detail, object shape and other appearance information. The paired depth image focuses on the 3D spatial representation of RGB image. They exist in different modalities. So two identical backbone networks, which are also called double-stream encoder, are used to extract the color and depth features, respectively. The backbone network employs ResNet-50 [73] which consists of a convolutional block *Conv*-1 and four residual blocks *Res*-$i(i=2,\cdots,5)$, and removes the last global pooling and fully connected layers, as shown in Table 1. The features of each block in RGB stream and depth stream are represented as side-output features $\{F_i^c|i=1,\cdots,5\}$ and $\left\{F_i^d|i=1,\cdots,5\right\}$, respectively.

In order to enhance the side-output features, CBAM attention modules [74] are inserted in each residual block. It assigns the attentive weight to the side-output feature, so that feature is enhanced in the positions which show higher response to salient objects and in the channels which exhibit the salient foreground regions. It can be described as:

$$\hat{F}_i^c = CBAM\left(F_i^c\right), \quad i=2,3,4,5 \tag{1}$$

$$\hat{F}_i^d = CBAM\left(F_i^d\right), \quad i=2,3,4,5 \tag{2}$$

where *CBAM* denotes the attention module in [74]. Note that the first block does not use CBAM for retaining more local detail.

3.2. Cross-modal and cross-level salient edge detector

The continuous convolutions with strides and pooling operations in the encoder make the feature in the deeper layer locate salient regions more accurately, but reduce the size of feature maps and lose the spatial structure, and deteriorate the edges of the salient objects. Although U-Net [7] and PFN[8] can progressively restore the size of feature map and refine the details of the salient object by upsampling the feature in the higher layer and short connection between encoder and decoder, the boundary of the salient object is still indistinct.

Considering that the low side-output feature preserves the better edge information [75,76,23] and the high side-output feature indicates the position of salient objects [24], they are combined to extract salient edge information, to make the boundary of salient object clearer. But different from EGNet [24] designed for

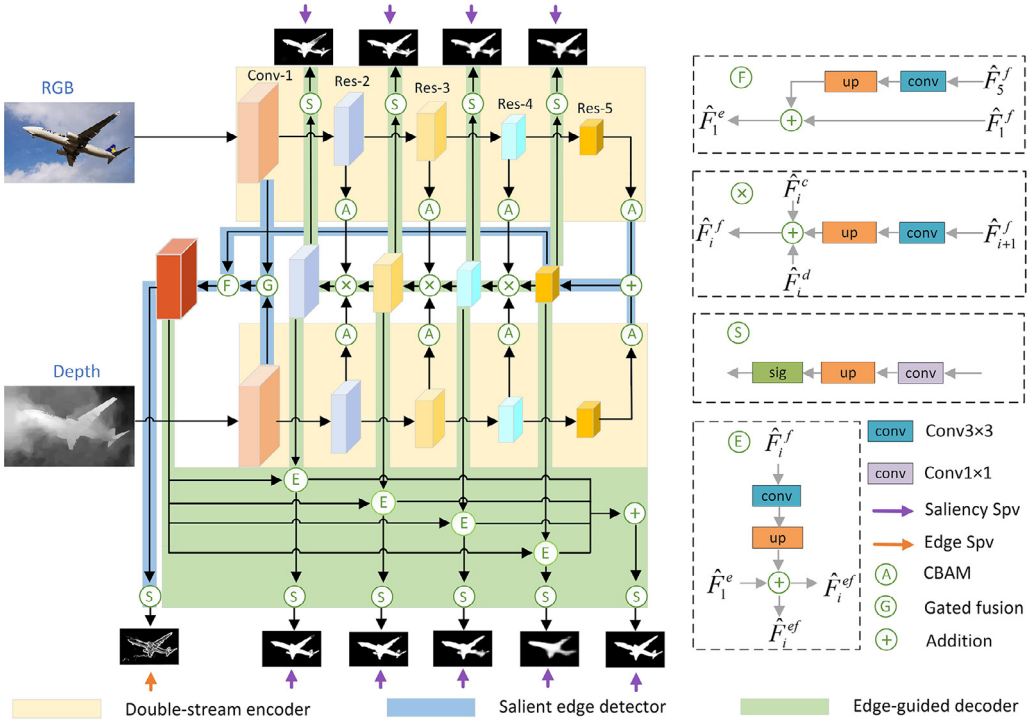


Fig. 3. The overall architecture of cross-modal edge-guided salient object detection in RGB-D images.

Table 1
Details of each block in the encoder, in which C×H×W denotes the channel, height and width of the feature map, respectively.

	Layer	Feature size
$Conv-1$	$64 \times 7 \times 7$, stride 2	64×112×112
	3×3 max pool, stride 2	256×56×56
$Res-2$	$\begin{bmatrix} 64 \times 1 \times 1 \\ 64 \times 3 \times 3 \\ 256 \times 1 \times 1 \\ CBAM \end{bmatrix} \times 3$, stride 2	256×56×56
$Res-3$	$\begin{bmatrix} 128 \times 1 \times 1 \\ 128 \times 3 \times 3 \\ 512 \times 1 \times 1 \\ CBAM \end{bmatrix} \times 4$, stride 2	512×28×28
$Res-4$	$\begin{bmatrix} 256 \times 1 \times 1 \\ 256 \times 3 \times 3 \\ 1024 \times 1 \times 1 \\ CBAM \end{bmatrix} \times 6$, stride 2	1024×14×14
$Res-5$	$\begin{bmatrix} 512 \times 1 \times 1 \\ 512 \times 3 \times 3 \\ 2048 \times 1 \times 1 \\ CBAM \end{bmatrix} \times 3$, stride 1, dilation 2	2048×14×14

RGB image salient object detection, we need to consider how to extract the better edge feature in cross-modal RGB-D image.

Since edge information has been diluted under the continuous convolutions, the feature in the highest layer show the very weak boundary information but the very accurate location information of the salient object. So the color and depth feature in the fifth layer are combined directly to express the position information. On the contrary, the feature in the lowest layer shows the more local details but a lot of noise, so a gated fusion layer [77,34] is used in the first layer to better fuse the color and depth information to prevent the influence of the poor depth image.

Therefore, we designed cross-modal cross-level salient edge detector, which obtains accurate salient edge features by fusing high-level information and low-level information from RGB and depth stream.

To be specific, high-level information uses attentive feature of the fifth block in both streams, and low-level information uses side-output feature of the first block in both streams.

At first, $\widehat{F}_5^c$ and $\widehat{F}_5^d$ are fused to generate $\widehat{F}_5^f$, which is denoted as:

$$\widehat{F}_5^f = \widehat{F}_5^c + \widehat{F}_5^d \tag{3}$$

where "+" denotes the element-wise addition operation.

Second, F_1^c and F_1^d are fused to generate $\widehat{F}_1^f$ by the gated fusion layer [77,34]. To be specific, as shown in Fig. 4 the feature F_1^c and F_1^d are concatenated and then performed a 1×1 convolution to obtain a 64-channel fused feature, and then performed the sigmoid function to generate the probability matrix $P \in \mathbb{R}^{64\times112\times112}$.

$$P = sig\left(Conv_{1\times1}\left(F_1^c \oplus F_1^d\right)\right) \tag{4}$$

where "$\oplus$" denotes the channel-wise concatenation operation, "sig" denotes sigmoid function, "$Conv_{1\times1}$" denotes convolution operation with 1×1 kernel. Let P and 1-P denote the weighted gates of the RGB stream and the depth stream, respectively. Then the gated fusion feature $\hat{F}_1^f$ is defined as:

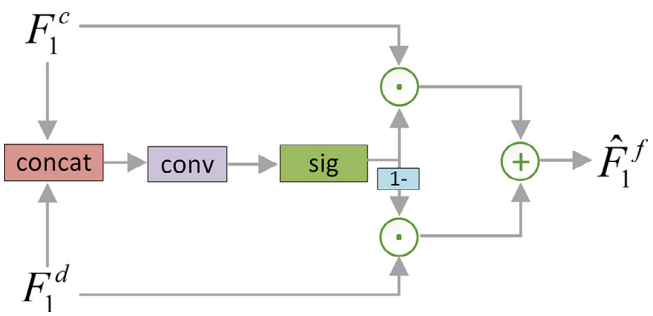


Fig. 4. The detail of the gated fusion layer.

$$\widehat{F}_1^f = P \odot F_1^c + (1-P) \odot F_1^d \tag{5}$$

"$\odot$" denotes Hadamard product.

At last, $\widehat{F}_5^f$ and $\widehat{F}_1^f$ are fused to generate salient edge feature $\hat{F}_1^e$. To be specific, the high-level fusion feature $\widehat{F}_5^f$ are firstly fed into a convolution layer, and then upsampling by bilinear interpolation operation, which retain the same spatial size and channel number as $\widehat{F}_1^f$, and last fused with $\widehat{F}_1^f$ by the element-wise addition operation to generate the salient edge feature $\hat{F}_1^e$. It can be denoted as:

$$\hat{F}_1^e = \hat{F}_1^f + Up\left(Conv_{3\times3}\left(\widehat{F}_5^f\right); \widehat{F}_1^f\right) \tag{6}$$

where $Up(\theta1;\theta2)$ is bilinear interpolation operation which aims to up-sample the feature $\theta1$ to the same size as the feature $\theta2$, $Conv_{3\times3}$ is convolution operation with 3×3 kernels.

3.3. Edge-guided cross-modal decoder

Double-stream encoder generates attentive side-output features and salient edge detector products salient edge feature. Next, in order to achieve pixel-level salient object detection task, edge-guided cross-modal decoder is designed to achieve the decoding process.

To be specific, decoder begins from the 5^{th} block, the fused feature $\widehat{F}_5^f$ is first upsampled to the same size as the feature in the 4^{th} block, and then combined with the attentive side-output features in the 4^{th} block from RGB and depth stream. Progressive restoration happens in the 3^{th} and 2^{th} blocks, and generates the fused feature $\widehat{F}_i^f$ $(i=4,3,2)$. They are denoted as:

$$\widehat{F}_i^f = Up\left(Conv_{3\times3}\left(\widehat{F}_{i+1}^f\right); \widehat{F}_i^c\right) + \widehat{F}_i^c + \widehat{F}_i^d, i \in \{4,3,2\} \tag{7}$$

These fused features are then performed convolution operation with 1×1 kernel, and then upsampled to the same size as the ground truth, and then performed the sigmoid function to generate saliency map $S_i^f (i=2,3,4,5)$. Note that the first block is not utilized in the decoding process.

$$S_i^f = sig\left(Up\left(Conv_{1\times1}\left(\widehat{F}_i^f\right); GT\right)\right), i \in \{2,3,4,5\} \tag{8}$$

where GT denotes ground truth saliency map.

Since the salient boundaries are diluted during the upsampling process, salient edge feature $\hat{F}_1^e$ is incorporated into the fused features to enhance the boundary information. To be specific, the fused feature $\hat{F}_i^f (i=2,3,4,5)$ need to be upsampled to the same size as the feature $\hat{F}_1^e$, and be combined with the edge features $\hat{F}_1^e$ to generate the edge-guided feature $\hat{F}_i^{ef} (i=2,3,4,5)$ by element-wise addition. The process can be described as:

$$\hat{F}_i^{ef} = \hat{F}_1^e + Up\left(Conv_{3\times3}\left(\hat{F}_i^f\right); \hat{F}_1^e\right), i \in \{2,3,4,5\} \tag{9}$$

The edge-guided feature $\hat{F}_i^{ef}$ is then fed into a convolution layer with 1×1 kernel, upsampling layer and the sigmoid function to generate edge-guided saliency map $S_i^e (i=2,3,4,5)$.

$$S_i^e = sig\left(Up\left(Conv_{1\times1}\left(\hat{F}_i^{ef}\right); GT\right)\right), i \in \{2,3,4,5\} \tag{10}$$

At the same time, salient edge feature $\hat{F}_1^e$ is performed the same operation to generate predicted salient edge map S_1^e.

$$S_1^e = sig\left(Up\left(Conv_{1\times1}\left(\hat{F}_1^e\right); GT\right)\right) \tag{11}$$

At last four edge-guided feature maps $\hat{F}_i^{ef} (i=2,3,4,5)$ are combined to generate a final saliency map S.

$$S = sig\left(Up\left(Conv_{1\times1}\left(\sum_{i=2}^{5}\hat{F}_i^{ef}\right); GT\right)\right) \tag{12}$$

3.4. Loss function

The loss functions L includes two parts: the loss of the edges L_e and the loss of saliency L_s. It is denoted as:

$$L = L_e + L_s \tag{13}$$

3.4.1. Edge loss

The edge ground truth can be easily got from saliency map ground truth by Canny edge detector [78]. It is used to supervise the salient edge map S_1^e. The loss of the edges L_e adopts the cross-entropy loss, and it is defined as:

$$L_e = -\sum_{j\in Z_+} logPr(y_j = 1|S_1^e) - \sum_{j\in Z_-} logPr(y_j = 0|S_1^e) \tag{14}$$

where Z_+ and Z_- denote the salient edge pixels set and background pixels set respectively. $Pr(y_j = 1|S_1^e)$ is the prediction map in which each value denotes the salient edge confidence for the pixel.

3.4.2. Saliency loss

There are nine saliency maps need to be supervised by the ground truth. They are four saliency maps $S_i^f (i=2,\cdots,5)$ from fused features, and four saliency maps $S_i^e (i=2,\cdots,5)$ from edge-guided saliency features, and a final saliency map S.

The loss of each saliency map is defined as:

$$L_s(P) = -\sum_{j\in Y_+} logPr(y_j = 1|P) - \sum_{j\in Y_-} logPr(y_j = 0|P) \tag{15}$$

where Y_+ and Y_- denote the salient region pixels set and non-salient pixels set respectively. $Pr(y_j = 1|P)$ is the prediction map P in which each value denotes the salient region confidence for the pixel.

Therefore saliency loss L_s is defined as:

$$L_s = \sum_{i=2}^{5} L_s\left(S_i^f\right) + \sum_{i=2}^{5} L_s\left(S_i^e\right) + L_s(S) \tag{16}$$

4. Experiments

4.1. Dataset

We conduct our experiments on seven benchmark datasets: NLPR [57] with 1000 images captured by Kinect, NJUD [58] with 2000 stereo images, STERE [79] with 1000 stereo images and DES [53] with 135 images captured by Kinect, SIP [33] with 1000 images captured by smart phone, LFSD[80] with 100 images captured by the Lytro camera and DUT-RGBD[81] with 1200 images captured by Lytro camera in real life scenes.

4.2. Evaluation metrics

PR curve [82], F-measure [83], mean absolute error (MAE) [84] and S-measure [85] are used to evaluate the performance of our model.

PR curve is plotted by setting a group of thresholds on the saliency maps to get the binary masks and further comparing them with the ground truth.

F-measure is the weighted harmonic mean of precision and recall, which can be defined as:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{17}$$

where the value of β^2 is 0.3 to emphasize the importance of accuracy [26]. It uses the adaptive threshold which is defined as twice the mean value of the saliency map to segment the saliency map to binary map and further compute the precision and recall.

MAE normalizes the saliency and ground-truth maps to [0,1], and calculates the mean absolute error between them, it is defined as:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |S(x,y) - Y(x,y)| \tag{18}$$

where W and H are the width and height of the feature map.

S-measure simultaneously evaluates region-aware and object-aware structural similarity between saliency map and ground truth, and it is defined as:

$$S_\lambda = \lambda * S_o + (1-\lambda) * S_r \tag{19}$$

where S_o and S_r are the object-aware and region-aware structural similarity, respectively, λ is the balance parameter and set as 0.5 in our experiment.

4.3. Implementation details

We use Pytorch [86] to implement our model, and the ResNet-50 [73] pre-trained on ImageNet [87] as our backbone network. During the inference stage, images are simply resized to 224×224, and then fed into the network to obtain prediction without any other post-processing. In the experiments we minimize objective function using Adaptive Moment Estimation (Adam) [88], with a batch size of 1, and the learning rate is initialized as 0.0005. The total training time is about 20 h. The inference of an image takes about 0.02s with a single NVIDIA GTX 1080Ti GPU.

4.4. Comparison with SOTA models

We compare the proposed model with 9 SOTA models, including AFNet [89], CTMF [90], MMCI [91], PCF [92], TANet [93], CPFP [94], DMRA [81], A2dele [95] and cmSalGAN [25]. For fair comparison, when comparing with DMRA [81] and A2dele [95], we use the same training datasets as DMRA [81] which contains 800 samples from the DUT-RGBD dataset [81], 1485 samples from NJUD [58] and 700 samples from NLPR [57]. When comparing with the others, we use the same training datasets as CTMF [90] which contains 1400 samples from the NJUD dataset [58] and 650 samples from NLPR [57].

Quantitative Comparison: As shown in Table 2, compared with other models, our model achieves the best scores on seven datasets in term of all the evaluation metrics under two training strategies. It proves the superior performance of our proposed model. we also find that the performance of our model is improved when adding 800 images in DUT-RGBD training dataset, especially in LFSD dataset and DUT-RGBD testing dataset. The images in these two datasets are both captured by Lytro camera. The train of the model on more training dataset enhances the detection ability in these light field images. In addition, Fig. 5 shows PR curve comparison among all the models. It can be seen that our model is always better than all other models under different thresholds, which means that our method has a good ability to detect salient objects in RGB-D image and generate accurate saliency maps.

Visual Performance: Fig. 6 shows the visual performance comparison. It can be seen that the proposed method can produce more accurate saliency maps with complete object outlines and sharp boundary details due to cross-modal edge-guided saliency detection. For example, our method can distinguish the salient objects with sharp boundary in complex environments in the 1^{th}-2^{th} rows. For the complex texture images in the 3^{th}-4^{th} rows, our method can depict the distinct texture boundary. For small object images in the 5^{th}-6^{th} rows and the multiple salient objects images in the 7^{th}-8^{th} rows, our method has the better performance too.

4.5. Ablation study

We conduct ablation studies on NLPR, NJUD and STERE datasets based on the same training datasets as CTMF[90] to investigate the contributions of different mechanisms in the proposed method. The baseline model used here contains a ResNet-50 backbone network equipped with CBAM. It takes RGB image and the paired depth image as input and adopts double stream U-Net structure. The performance of baseline model without any additional mechanisms is illustrated in Table 3 No.1. Based on the baseline model,

Table 2
S-measure, adaptive F-measure, MAE comparisons with SOTA models on different datasets. The models marked with * are trained on NJUD + NLPR + DUT-RGBD, the rest models are trained on NJUD + NLPR. The best results are in bold.

Datasets	Metric	AFNet18	CTMF18	MMCI18	PCF18	TANet19	CPFP19	cmSalGAN20	Ours	DMRA19*	A2dele20*	Ours*
NLPR	S↑	.799	.860	.856	.874	.886	.888	.922	**.925**	.899	.898	**.929**
	adpF$_\beta$ ↑	.747	.724	.730	.795	.796	.823	.863	**.882**	.854	.874	**.875**
	MAE↓	.058	.056	.059	.044	.041	.036	.027	**.025**	.031	.028	**.025**
NJUD	S↑	.772	.849	.858	.877	.878	.879	.903	**.914**	.886	.871	**.908**
	adpF$_\beta$↑	.768	.788	.812	.844	.844	.837	.874	**.894**	.872	.872	**.892**
	MAE↓	.100	.085	.079	.059	.060	.053	.046	**.040**	.051	.051	**.039**
STERE	S↑	.825	.848	.873	.875	.871	.879	.896	**.902**	.886	.878	**.895**
	adpF$_\beta$↑	.807	.771	.829	.826	.835	.830	.863	**.876**	.844	.870	**.886**
	MAE↓	.075	.086	.068	.064	.060	.051	.050	**.046**	.047	.047	**.043**
DES	S↑	.770	.863	.848	.842	.858	.872	.913	**.928**	.901	.886	**.923**
	adpF$_\beta$↑	.730	.778	.762	.782	.795	.829	.869	**.892**	.866	.866	**.896**
	MAE↓	.068	.055	.065	.049	.046	.038	.028	**.023**	.029	.028	**.022**
SIP	S↑	.720	.716	.833	.842	.835	.850	.865	**.876**	.806	.828	**.867**
	adpF$_\beta$↑	.705	.684	.795	.825	.809	.819	.844	**.848**	.819	.827	**.863**
	MAE↓	.118	.139	.086	.071	.075	.064	.064	**.056**	.085	.070	**.055**
LFSD	S↑	.738	.796	.787	.794	.801	.828	.830	**.855**	.847	.833	**.866**
	adpF$_\beta$↑	.742	.782	.779	.792	.794	.813	.831	**.849**	.849	.831	**.862**
	MAE↓	.133	.119	.132	.112	.111	.088	.097	**.075**	.075	.077	**.063**
DUT-RGBD	S↑	.702	.831	.791	.801	.808	.818	.867	**.883**	.888	.885	**.924**
	adpF$_\beta$↑	.743	.792	.753	.760	.779	.783	.844	**.864**	.883	.892	**.922**
	MAE↓	.122	.097	.113	.100	.093	.076	.067	**.058**	.048	.042	**.031**

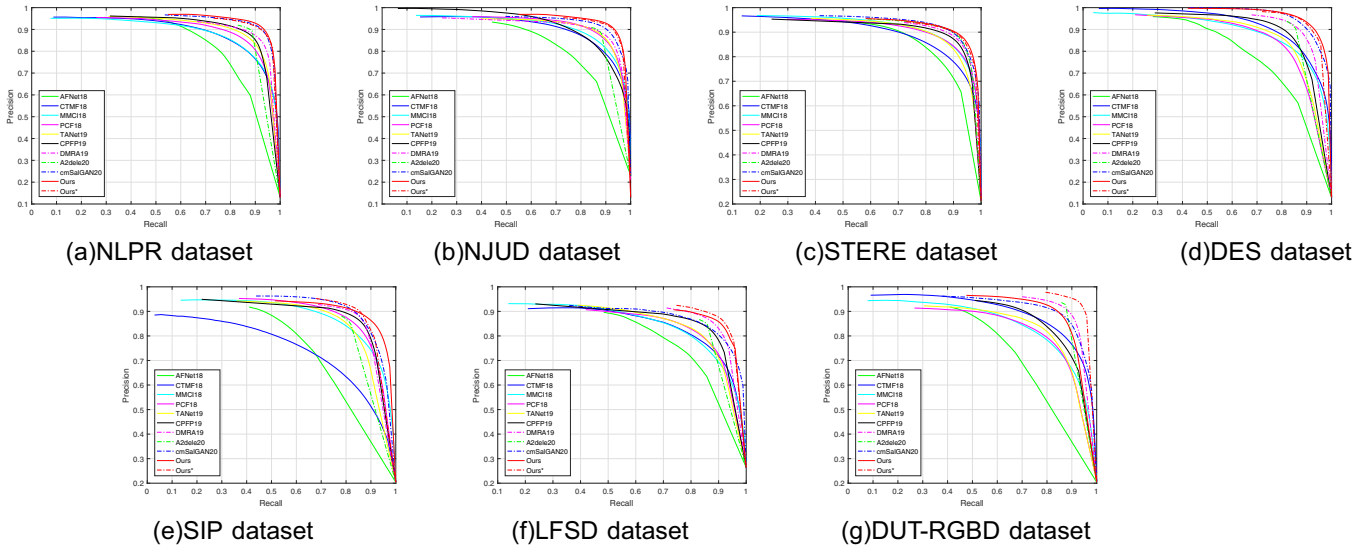


Fig. 5. P-R curves comparison with the state-of-the-art methods.

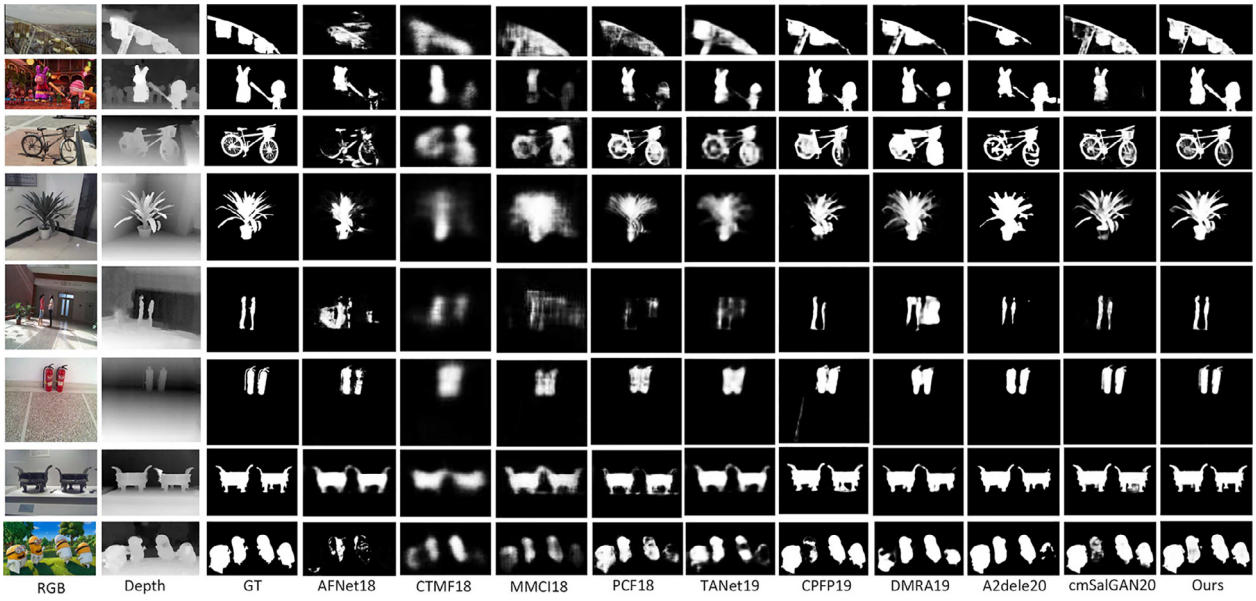


Fig. 6. Visual comparison of the proposed model with SOTA methods. Apparently, saliency maps produced by our model are clearer and more accurate than others and our results are more consistent with the ground truths.

Table 3
Ablation experiments of different components. The best result is in bold.

Variant	Candidate			NLPR			NJUD			STERE		
	Baseline	EG	GF	S↑	F_β ↑	MAE↓	S↑	F_β↑	MAE↓	S↑	F_β↑	MAE↓
No.1	✓			.910	.859	.031	.891	.872	.048	.885	.869	.051
No.2	✓	✓		.923	.870	.027	.912	.884	.042	.901	.870	.049
No.3	✓	✓	✓	**.925**	**.882**	**.025**	**.914**	**.894**	**.040**	**.902**	**.876**	**.046**

we gradually add different mechanisms and test various combinations. These candidates are edge-guided decoder (EG), gated fusion edge extractor(GF). In Table 3 No.2, by applying EG, the performance is boosted greatly. It benefits from refining the contour detail of salient object by extracting salient edge feature and integrating edge feature into decoding process. We can see that EG increases S-measure by 1.9% and F-measure by 0.9%, and improves MAE by 9.8% in average compared with No.1. In Table 3 No.3, by applying GF in the first layer rather than element-wise addition, the performance is improved to some extent. Through the adaptive gated fusion, the contribution of RGB feature and depth feature in the first block is weighted, better edge feature is generated. We can see that GF increases S-measure by 0.2% and F-measure by 1.1%, and improves MAE by 6.1% in average compared with No.2. The

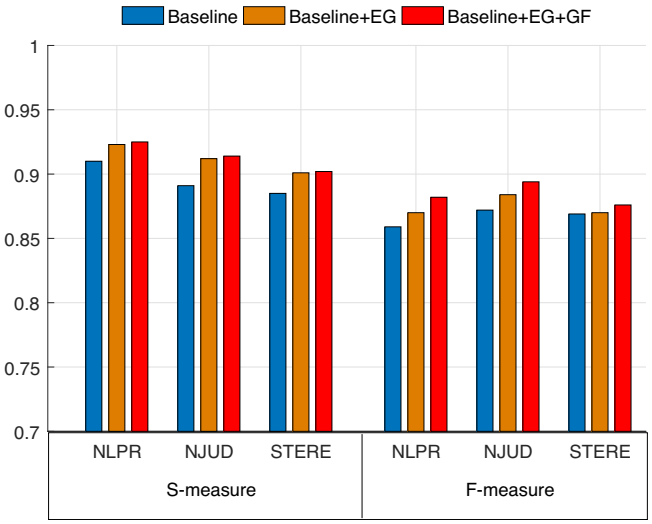


Fig. 7. Comparisons of S-measure and F-measure to evaluate the contribution of EG and GF components. Baseline + EG represents our baseline model with edge-guided decoding process. Baseline + EG + GF represents our baseline model with edge-guided decoding process and the gated fusion layer in the first layer. The comparisons are evaluated on NLPR, NJUD and STERE datasets respectively.

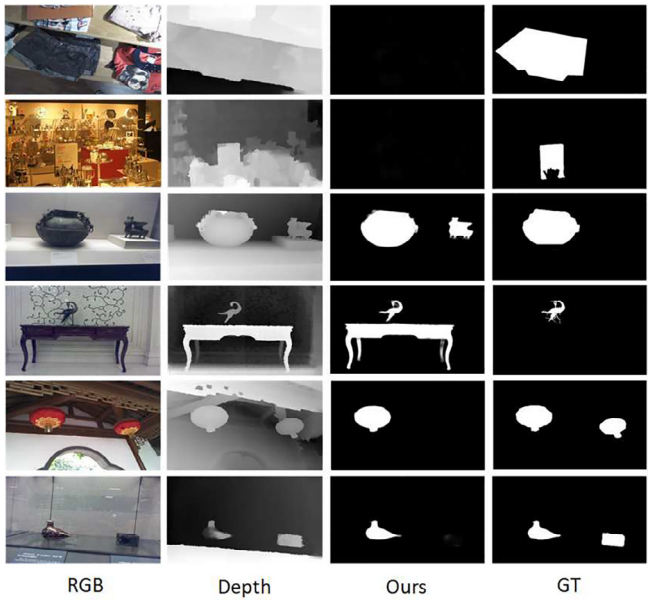


Fig. 8. Failure Cases.

Table 4
The setting of gated fusion in different layers. The best results are in bold. GF_i represents the gated fusion layer which is added in the i^{th} block.

Variant	Candidate			NLPR			NJUD			STERE		
	Baseline + EG	GF_1	GF_5	S↑	F_β ↑	MAE↓	S↑	F_β↑	MAE↓	S↑	F_β↑	MAE↓
No.1	✓			.923	.870	.027	.912	.884	.042	.901	.870	.049
No.2	✓	✓		**.925**	**.882**	**.025**	**.914**	**.894**	**.040**	**.902**	**.876**	**.046**
No.3	✓		✓	.921	.866	.027	.908	.882	.043	.897	.853	.048
No.4	✓	✓	✓	.920	.869	.028	.906	.881	.044	.895	.852	.047

effectiveness of EG and GF components is clearer demonstrated in Fig. 7. S-measure and F-measure are all improved by gradually adding EG and GF mechanisms.

Further, we also demonstrate the influence of the position of the gated fusion layer. Since edge feature extractor uses the first and the fifth block of the backbone network, there are four variants about the position of the gated fusion layer. Ablation study in Table 4 shows that when adding the gated fusion layer in the first block, the results are the best. It also verifies that depth features in the shallow layer may have some noise although it has more spatial detail. The performance is improved by reducing the influence of poor depth image in the shallow layer based on the gated fusion.

4.6. Failure cases

The proposed method has the good detection performance in most cases. But when there is a complex background in the color image and depth image is not enough to supplement it, failure cases will occur, which is shown in Fig. 8. In the first two rows, some images are considered non-salient images, and no salient objects are detected. In the middle two rows, some non-salient regions are regarded as salient objects. In the last two rows, some salient regions are omitted. It is a big challenge for future research too.

5. Conclusion

In this paper, we propose a cross-modal edge-guided network for RGB-D image salient object detection. The network obtains more accurate edge information by fusing RGB and depth information, which is then fed back to guide the salient features to reinforce the sharp boundary. In the process of extracting edge information, the gated fusion layer is introduced to reduce the influence of poor depth image. Experimental results show the effect of the edge guidance and gated fusion, and the whole model outperforms the latest salient object detection methods.

CRediT authorship contribution statement

Zhengyi Liu: Methodology, Writing - review & editing. **Kaixun Wang:** Writing - original draft. **Hao Dong:** Visualization, Validation. **Yuan Wang:** Data curation.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

We thank all anonymous reviewers for their valuable comments. This research is supported by Natural Science Foundation of Anhui Province (1908085MF182) and Key Program of Natural Science Project of Educational Commission of Anhui Province (KJ2019A0034).

References

[1] Q. Hou, P. Jiang, Y. Wei, M.-M. Cheng, Self-erasing network for integral object attention, Advances in Neural Information Processing Systems (2018) 549–559.

[2] Z. Ren, S. Gao, L.-T. Chia, I.W.-H. Tsang, Region-based saliency detection and its application in object recognition, IEEE Trans. Circuits Syst. Video Technol. 24 (5) (2013) 769–779.

[3] A. Das, H. Agrawal, L. Zitnick, D. Parikh, D. Batra, Human attention in visual question answering: Do humans and deep networks look at the same regions?, Comput. Vis. Image Underst. 163 (2017) 90–100.

[4] S. Hong, T. You, S. Kwak, B. Han, Online tracking by learning discriminative saliency map with convolutional neural network, in: International conference on machine learning, 2015, pp. 597–606.

[5] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556..

[6] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 3431–3440.

[7] O. Ronneberger, P. Fischer, T. Brox, U-net: Convolutional networks for biomedical image segmentation, in: International Conference on Medical image computing and computer-assisted intervention, Springer, 2015, pp. 234–241.

[8] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2117–2125.

[9] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, P.-M. Jodoin, Non-local deep features for salient object detection, in: Proceedings of the IEEE Conference on computer vision and pattern recognition, 2017, pp. 6609–6617.

[10] Z. Chen, H. Zhou, X. Xie, J. Lai, Contour loss: Boundary-aware learning for salient object segmentation, arXiv preprint arXiv:1908.01975..

[11] Y. Wang, X. Zhao, X. Hu, Y. Li, K. Huang, Focal boundary guided salient object detection, IEEE Trans. Image Process. 28 (6) (2019) 2813–2824.

[12] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, J. Jiang, A simple pooling-based design for real-time salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 3917–3926.

[13] 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.

[14] W. Guan, T. Wang, J. Qi, L. Zhang, H. Lu, Edge-aware convolution neural network based salient object detection, IEEE Signal Process. Lett. 26 (1) (2018) 114–118.

[15] Y. Zhuge, G. Yang, P. Zhang, H. Lu, Boundary-guided feature aggregation network for salient object detection, IEEE Signal Process. Lett. 25 (12) (2018) 1800–1804.

[16] 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.

[17] S. Zhou, J. Zhang, J. Wang, F. Wang, D. Huang, SE2Net: Siamese edge-enhancement network for salient object detection, arXiv preprint arXiv:1904.00048..

[18] J. Su, J. Li, Y. Zhang, C. Xia, Y. Tian, Selectivity or invariance: Boundary-aware salient object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 3799–3808.

[19] F. Lin, C. Yang, H. Li, B. Jiang, Boundary-aware salient object detection via recurrent two-stream guided refinement network, arXiv preprint arXiv:1912.05236..

[20] J. Zhang, Y. Dai, F. Porikli, M. He, Deep edge-aware saliency detection, arXiv preprint arXiv:1708.04366..

[21] W. Wang, S. Zhao, J. Shen, S.C. Hoi, A. Borji, Salient object detection with pyramid attention and salient edges, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2019, pp. 1448–1457.

[22] 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.

[23] P. Zhang, D. Wang, H. Lu, H. Wang, X. Ruan, Amulet: Aggregating multi-level convolutional features for salient object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 202–211.

[24] J.-X. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, M.-M. Cheng, EGNet: Edge guidance network for salient object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8779–8788.

[25] B. Jiang, Z. Zhou, X. Wang, J. Tang, B. Luo, cmSalGAN:RGB-D salient object detection with cross-view generative adversarial networks, IEEE Trans. Multimedia (2020) 1–10.

[26] 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.

[27] Z. Liu, Q. Li, W. Li, Deep layer guided network for salient object detection, Neurocomputing 372 (2020) 55–63.

[28] C. Fosco, A. Newman, P. Sukhum, Y.B. Zhang, N. Zhao, A. Oliva, Z. Bylinskii, How much time do you have? modeling multi-duration saliency, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4473–4482.

[29] S.-A. Rebuffi, R. Fong, X. Ji, A. Vedaldi, There and back again: Revisiting backpropagation saliency methods, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8839–8848.

[30] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, L. Yang, , Interactive two-stream decoder for accurate and fast saliency detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9141–9150.

[31] 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.

[32] 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.

[33] 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 Networks Learn. Syst. (2020) 1–15.

[34] Z. Liu, W. Zhang, P. Zhao, A cross-modal adaptive gated fusion generative adversarial network for rgb-d salient object detection, Neurocomputing 387 (2020) 210–220.

[35] J. Zhang, D.-P. Fan, Y. Dai, S. Anwar, F.S. Saleh, T. Zhang, N. Barnes, Uc-net: uncertainty inspired rgb-d saliency detection via conditional variational autoencoders, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 8582–8591.

[36] 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.

[37] 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.

[38] L. Wei, S. Zhao, O.E.F. Bourahla, X. Li, F. Wu, Y. Zhuang, Deep group-wise fully convolutional network for co-saliency detection with graph propagation, IEEE Trans. Image Process. 28 (10) (2019) 5052–5063.

[39] K. Zhang, T. Li, B. Liu, Q. Liu, Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing, IEEE CVPR (2019) 3095–3104.

[40] C. Wang, Z.-J. Zha, D. Liu, H. Xie, Robust deep co-saliency detection with group semantic, AAAI (2019) 8917–8924.

[41] R. Huang, W. Feng, Z. Wang, Y. Xing, Y. Zou, Exemplar-based image saliency and co-saliency detection, Neurocomputing 371 (2020) 147–157.

[42] D.-P. Fan, Z. Lin, G.-P. Ji, D. Zhang, H. Fu, M.-M. Cheng, Taking a deeper look at co-salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 2919–2929.

[43] K. Zhang, T. Li, S. Shen, B. Liu, J. Chen, Q. Liu, Adaptive graph convolutional network with attention graph clustering for co-saliency detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9050–9059.

[44] D.-P. Fan, W. Wang, M.-M. Cheng, J. Shen, Shifting more attention to video salient object detection, IEEE CVPR (2019) 8554–8564.

[45] J. Shen, J. Peng, L. Shao, Submodular trajectories for better motion segmentation in videos, IEEE TIP 27 (6) (2018) 2688–2700.

[46] W. Wang, J. Shen, F. Guo, M.-M. Cheng, A. Borji, Revisiting video saliency: A large-scale benchmark and a new model, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4894–4903.

[47] A. Tsiami, P. Koutras, P. Maragos, Stavis: Spatio-temporal audiovisual saliency network, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 4766–4776.

[48] T. Wang, Y. Piao, X. Li, L. Zhang, H. Lu, Deep learning for light field saliency detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 8838–8848.

[49] Y. Piao, Z. Rong, M. Zhang, H. Lu, Exploit and replace: An asymmetrical two-stream architecture for versatile light field saliency detection., AAAI (2020) 11865–11873.

[50] Y. Zeng, P. Zhang, J. Zhang, Z. Lin, H. Lu, Towards high-resolution salient object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2019, pp. 7234–7243.

[51] C. Li, R. Cong, J. Hou, S. Zhang, Y. Qian, S. Kwong, Nested network with two-stream pyramid for salient object detection in optical remote sensing images, IEEE Trans. Geosci. Remote Sens. 57 (11) (2019) 9156–9166.

[52] K. Desingh, K.M. Krishna, D. Rajan, C. Jawahar, Depth really matters: Improving visual salient region detection with depth, in: BMVC, 2013, pp. 1–11..

[53] 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.

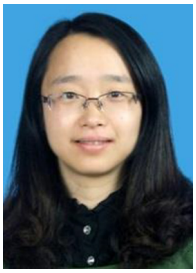
[54] J. Ren, X. Gong, L. Yu, W. Zhou, M. Ying Yang, Exploiting global priors for RGB-D saliency detection, IEEE CVPR (2015) 25–32.

[55] J. Guo, T. Ren, J. Bei, Y. Zhu, Salient object detection in RGB-D image based on saliency fusion and propagation, in: Proceedings of the 7th International Conference on Internet Multimedia Computing and Service, 2015, pp. 1–5.

[56] Y. Tang, R. Tong, M. Tang, Y. Zhang, Depth incorporating with color improves salient object detection, TVC 32 (1) (2016) 111–121.

[57] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, RGB-D salient object detection: a benchmark and algorithms, in: ECCV, Springer, 2014, pp. 92–109..

[58] R. Ju, L. Ge, W. Geng, T. Ren, G. Wu, Depth saliency based on anisotropic center-surround difference, IEEE ICIP, IEEE (2014) 1115–1119.
[59] D. Feng, N. Barnes, S. You, C. McCarthy, Local background enclosure for RGB-D salient object detection, IEEE CVPR (2016) 2343–2350.
[60] D. Feng, N. Barnes, S. You, HOSO: Histogram of surface orientation for RGB-D salient object detection, in: Digital Image Computing: Techniques and Applications (DICTA), IEEE, 2017, pp. 1–8..
[61] F. Liang, L. Duan, W. Ma, Y. Qiao, Z. Cai, L. Qing, Stereoscopic saliency model using contrast and depth-guided-background prior, Neurocomputing 275 (2018) 2227–2238.
[62] C. Zhu, G. Li, W. Wang, R. Wang, An innovative salient object detection using center-dark channel prior, in: Proceedings of the IEEE International Conference on Computer Vision Workshops, 2017, pp. 1509–1515.
[63] H. Xue, Y. Gu, Y. Li, J. Yang, RGB-D saliency detection via mutual guided manifold ranking, IEEE ICIP, IEEE (2015) 666–670.
[64] J. Guo, T. Ren, J. Bei, Salient object detection for RGB-D image via saliency evolution, IEEE ICME, IEEE (2016) 1–6.
[65] 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 TIP 26 (9) (2017) 4204–4216.
[66] P. Hu, B. Shuai, J. Liu, G. Wang, Deep level sets for salient object detection, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 2300–2309.
[67] 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 the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3203–3212.
[68] G. Li, Y. Yu, Deep contrast learning for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 478–487.
[69] N. Liu, J. Han, M.-H. Yang, Picanet: Learning pixel-wise contextual attention for saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3089–3098.
[70] S. Xie, Z. Tu, Holistically-nested edge detection, in: Proceedings of the IEEE international conference on computer vision, 2015, pp. 1395–1403.
[71] Z. Tu, Y. Ma, C. Li, J. Tang, B. Luo, Edge-guided non-local fully convolutional network for salient object detection, IEEE Trans. Circuits Syst. Video Technol. (2020) 1–10.
[72] C.L. Zitnick, P. Dollár, Edge boxes: Locating object proposals from edges, in: European conference on computer vision, Springer, 2014, pp. 391–405..
[73] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 770–778.
[74] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, CBAM: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19..
[75] M.D. Zeiler, R. Fergus, Visualizing and understanding convolutional networks, in: European conference on computer vision, Springer, 2014, pp. 818–833.
[76] A. Mahendran, A. Vedaldi, Understanding deep image representations by inverting them, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2015, pp. 5188–5196.
[77] Y. Cheng, R. Cai, Z. Li, X. Zhao, K. Huang, Locality-sensitive deconvolution networks with gated fusion for RGB-D indoor semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 3029–3037.
[78] J. Canny, A computational approach to edge detection, IEEE Trans. Pattern Anal. Mach. Intell. 8 (6) (1986) 679–698.
[79] Y. Niu, Y. Geng, X. Li, F. Liu, Leveraging stereopsis for saliency analysis, IEEE CVPR, IEEE (2012) 454–461.
[80] N. Li, J. Ye, Y. Ji, H. Ling, J. Yu, Saliency detection on light field, IEEE CVPR (2014) 2806–2813.
[81] 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.
[82] A. Borji, M.-M. Cheng, H. Jiang, J. Li, Salient object detection: A benchmark, IEEE Trans. Image Process. 24 (12) (2015) 5706–5722.
[83] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: Computer vision and pattern recognition, 2009. cvpr 2009. ieee conference on, IEEE, 2009, pp. 1597–1604..
[84] F. Perazzi, P. Krähenbühl, Y. Pritch, A. Hornung, Saliency filters: Contrast based filtering for salient region detection, in: 2012 IEEE conference on computer vision and pattern recognition, IEEE, 2012, pp. 733–740..
[85] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, A. Borji, Structure-measure: A new way to evaluate foreground maps, in: IEEE ICCV, 2017, pp. 4558–4567..
[86] N. Ketkar, Introduction to pytorch, in: Deep learning with python, Springer, 2017, pp. 195–208..
[87] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, IEEE CVPR, Ieee (2009) 248–255.
[88] D.P. Kingma, J. Ba, Adam: A method for stochastic optimization, arXiv preprint arXiv:1412.6980..
[89] N. Wang, X. Gong, Adaptive fusion for RGB-D salient object detection, IEEE Access 7 (2019) 55277–55284.
[90] 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.
[91] 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 Recogn. 86 (2019) 376–385.
[92] H. Chen, Y. Li, Progressively complementarity-aware fusion network for RGB-D salient object detection, IEEE CVPR (2018) 3051–3060.
[93] H. Chen, Y. Li, Three-stream attention-aware network for rgb-d salient object detection, IEEE Trans. Image Process. 28 (6) (2019) 2825–2835.
[94] 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.
[95] Y. Piao, Z. Rong, M. Zhang, W. Ren, H. Lu, A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 9060–9069.



Zhengyi Liu is a professor in School of Computer Science and Technology, Anhui University, China. She received her B.S., M.S., and Ph.D. from Anhui University, China in 2001, 2004 and 2007, respectively. Her research interests include image and video processing, computer vision and deep learning.

KaiXun Wang is a M.S. Candidate of Anhui University. He received his B.S. from Jiangsu Ocean University, China in 2018. His research interests include image and video processing and computer vision.

Hao Dong is a M.S. Candidate of Anhui University. He received his B.S. from Jiang huai College of Anhui University, China in 2019. His research interests include image and video processing and computer vision.

Yuan Wang is a M.S. Candidate of Anhui University. He received his B.S. from University of Hefei, China in 2018. His research interests include image and video processing and computer vision.