Neurocomputing 387 (2020) 210–220

Contents lists available at ScienceDirect

Neurocomputing

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

A cross-modal adaptive gated fusion generative adversarial network for RGB-D salient object detection

Zhengyi Liu*, Wei Zhang, Peng Zhao

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 30 July 2019
Revised 15 December 2019
Accepted 11 January 2020
Available online 15 January 2020

Communicated by Dr. Yongsheng Dong

Keywords:
RGB-D salient object detection
Generative adversarial network
Cross-modal guidance
Adaptive gated fusion

ABSTRACT

Salient object detection in RGB-D images aims to identify the most attractive objects in a pair of color and depth images for the observer. As an important branch of salient object detection, it focuses on solving the following two major challenges: how to achieve cross-modal fusion that is efficient and beneficial for salient object detection; how to effectively extract the information of depth image with relatively poor quality. This paper proposes a cross-modal adaptive gated fusion generative adversarial network for RGB-D salient object detection by using color and depth images. Specifically, the generator network adopts double-stream encoder-decoder network and receives RGB and depth images at the same time. The proposed depthwise separable residual convolution module is used to deal with deep semantic information, and the processed feature is combined with side-output features of the encoder network progressively. In order to compensate for the shortcoming of poor quality of the depth image, the proposed method adds the cross-modal guidance from the side-output features of the RGB stream to the decoder network of depth stream. The discriminator network adaptively fuses the features of double streams using a gated fusion module, then sends the gated fusion saliency map to the discriminator to distinguish the similarity from ground-truth map. Adversarial learning forms the better generator network and discriminator network, and the gated fusion saliency map generated by the best generator network is served as final result. Experiments on five publicly RGB-D datasets demonstrate the effect of cross-modal fusion, depthwise separable residual convolution and adaptive gated fusion. Compared with the state-of-the-art methods, our method achieves the better performance.

1. Introduction

Salient object detection (SOD) is used to identify the most attractive objects of the image. It is a pre-processing process for most computer vision applications, such as image segmentation [1], face recognition [2], object detection [3], visual tracking [4], image retrieval [5], etc. At present, most of SOD methods adopt convolutional neural networks (CNNs) [6] which are generally superior to traditional methods [7] in the performance, and have achieved unprecedented test results on several challenging datasets. However, it is difficult to achieve good results only using RGB images in the scene with cluttered backgrounds, severe object occlusions and varying illuminations. But recently due to the advent of depth cameras, e.g. Kinect, high-quality synchronized visual cues (RGB data) and geometrical cues (depth data) can be captured to depict one scene. There is an opportunity to improve the performance of SOD by taking full advantage of two complementary modalities. Traditional methods simply concatenate the handcrafted RGB and depth features to represent each pixel or superpixel. Existing fusion methods based on convolutional neural networks are mainly divided into three types, as shown in Fig. 1. They are early fusion in which RGB and depth images are integrated into CNN [8] as a whole, middle fusion in which RGB and depth images are sent to different streams and the features of two modalities are fused at different scales [9] and late fusion in which RGB and depth images are sent to the double-stream network and fused in the final stage [10]. Therefore, how to achieve cross-modal fusion which is efficient and beneficial for salient object detection is our first task.

Meanwhile, encoder–decoder framework is often adopted in salient object detection network for dense pixel prediction. The feature of deep layers can locate the position of salient objects, and the feature of shallow layers can highlight the contours of the object. The features of deep layers in the encoder–decoder framework need to be upsampled progressively, which causes serious

* Corresponding author.
E-mail addresses: liuzywen@ahu.edu.cn (Z. Liu), 2446777351@qq.com (W. Zhang), 18868519@qq.com (P. Zhao).

https://doi.org/10.1016/j.neucom.2020.01.045
0925-2312/© 2020 Elsevier B.V. All rights reserved.

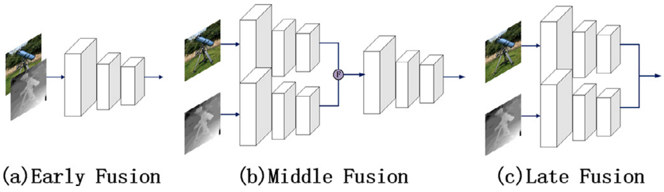


Fig. 1. Three different fusion manners.

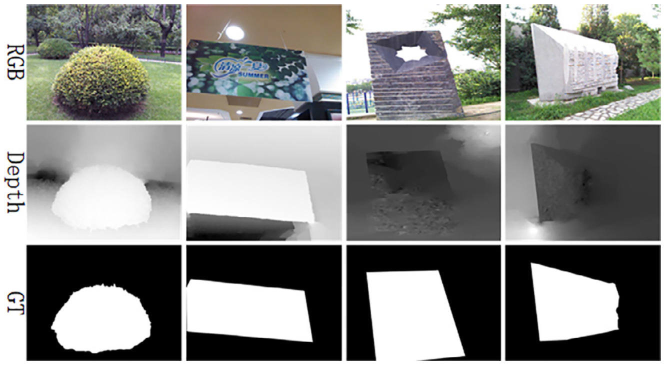


Fig. 2. Examples for effectiveness of depth maps in salient object detection. The depth maps in the first two columns can provide important spatial information for SOD, but the depth maps in the last two columns are less helpful for salient object detection.

information loss. Therefore, how to retain more semantic information in the upsampling process is our second task.

In addition, researches show that depth information can provide important spatial information for SOD, but depth maps with low quality are less helpful for the performance of SOD, as shown in the last two columns of Fig. 2. Therefore, how to effectively extract beneficial feature from depth image is our third task.

In the paper, we propose a cross-modal adaptive gated fusion generative adversarial network for RGB-D salient object detection, which mainly adopts generative adversarial network. The generator network(G-Net) is responsible for generating saliency map, and discriminator network(D-Net) takes part in judging whether the generated saliency map is close to ground-truth map. G-Net adopts double-stream framework, the proposed depthwise separable residual convolution module (DSRCM) is applied in each decoder subnetwork by combining deep layer semantic feature and shallow spatial feature progressively to generate the saliency map with more distinct boundaries. In order to achieve cross-modal fusion and remedy the drawback of the depth image with poor quality, the side-output features of the RGB stream are added to the depth stream for guiding the learning of depth stream. D-Net receives the features from the last layers of the double-stream network as input, and fuses the features of two different modalities by the adaptive gated fusion module to get gated fusion saliency map. Then, RGB image and gated fusion saliency map(RGB+S_{gated}) are delivered to discriminator to discriminate the similarity with the RGB image and ground-truth map(RGB+GT). G-Net and D-Net constitute the adversarial game which can enhance the ability of G-Net for generating saliency images close to ground-truth maps and the distinguishing ability of D-Net. By adversarial learning, G-Net becomes better and better, and its generated saliency maps are so good that they can not be distinguished from ground-truth map by D-Net. At last the gated fusion saliency map from G-Net is served as final result.

Our main contributions can be summarized as follows:

- A cross-modal adaptive gated fusion generative adversarial network is proposed for salient object detection in RGB-D images. The generator network uses the side-output features of the RGB stream to guide the learning of the depth stream. It compensates for the shortcoming that the feature of depth map is not clear and the reliability is not high.
- The depthwise separable residual convolution module is proposed to upsample deep semantic information of the double-stream network. It consists of residual convolutions in which standard convolution is replaced with more efficient depthwise separable convolution. So the deep semantic information is more retained and transited to shallow layers with lower computation cost.
- An adaptive gated fusion is used as a part of discriminator network to adaptively fuse the features of the RGB and depth streams. It can choose the best gated fusion saliency map to input to the discriminator for adversarial learning, and further improve the effect of the generator network.

2. Related work

2.1. RGB salient object detection

Traditional salient object detection methods rely heavily on handcraft features. For example, Li et al. [11] presents a saliency transfer method based on low-level handcraft features that involves the transfer of annotations from an example image to an input image. Most of the current salient object detection tasks adopt deep learning framework. Li and Yu [12] proposes an end-to-end deep contrast network for salient object detection, which contains a pixel-level fully convolutional stream and a segment-wise spatial pooling stream. The first stream directly produces a saliency map with pixel-level accuracy from an input image. The second stream extracts segment-wise features efficiently, and better models saliency discontinuities along object boundaries. Luo et al. [13] proposes a convolutional neural network that combines local and global information through a multi-resolution grid structure for salient object detection. Xie and Tu [14] develops a holistically-nested edge detection (HED) that emphasizes holistic image prediction and multi-scale feature learning. Hou et al. [15] introduces short connections to the skip-layer structures within the HED architecture to take full advantage of multi-scale features. Li et al. [16] proposes a multi-scale cascade network to identify the most attractive objects in an image. It can encode more global contextual information and obtain the saliency prior knowledge in the intermediate cascade stages. Li et al. [17] grafted salient object detection decoder onto the existing contour detection network to form a multi-task network architecture without using any manually labeled salient object masks. Han et al. [18] proposes a pixel-by-pixel method to add an edge convolution constraint to the U-Net to get the saliency map, which can fuse the features of different layers to reduce the loss of information. Li et al. [19] proposes a manifold matting framework for image matting. It apply some manifold learning methods in the framework to obtain several image matting methods, which provided a new view for us to process the RGB image in the future.

2.2. RGB-D salient object detection

RGB-D salient object detection introduces depth information on the basis of RGB salient object detection, so the processing of depth information becomes a key step. Ren et al. [20] proposes a two-stage RGB-D salient object detection framework that combines regional contrast with background, depth and direction priors to generate the saliency map. Han et al. [21] designs a convolutional neural network which transfers the structure of the RGB-based deep neural network to be applicable for depth view and fuses the deep representations of both views automatically

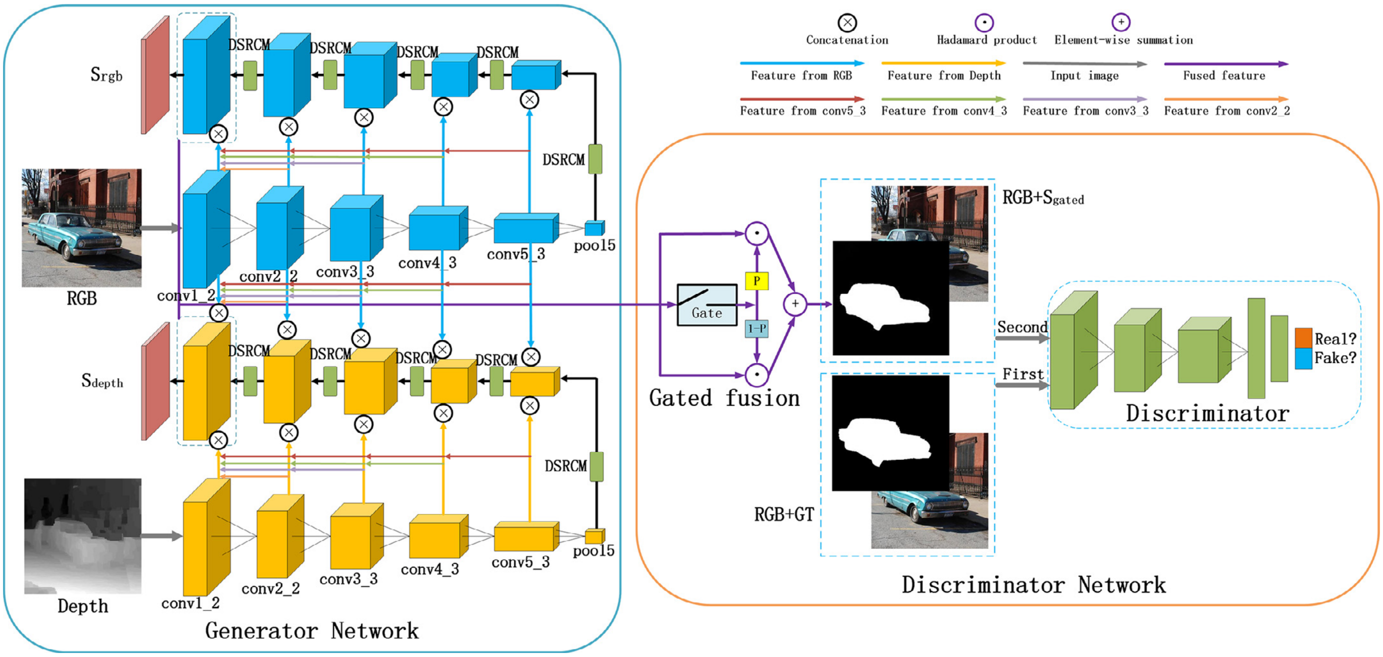


Fig. 3. The overall architecture of the cross-modal adaptive gated fusion generative adversarial network for salient object detection in RGB-D images.

to obtain the final saliency map. Chen and Li [22] proposes a complementarity-aware fusion(CA-Fuse) module in the convolutional neural network, and learns complementary information by introducing cross-modal residual functions and complementarity-aware supervision in each CA-Fuse module. Zhao et al. [23] proposes to use contrast prior to enhance depth information and process multi-scale cross-modal features through a fluid pyramid. The RGB-D fusion methods design complex structures, but mainly rely on simple equal weights fusion of feature connections and element addition, multiplication of prediction results. Moreover, the information of the depth map is not well processed, the useful information of the depth map is not fully utilized and the noise is not reasonably discarded. In contrast, our model has a better performance in this regard.

2.3. Generative adversarial network

Goodfellow et al. [24] first proposes generative adversarial network (GAN). It includes a generator network (GNet) and a discriminator network (DNet). In the training process, the goal of GNet is to generate a fake image to deceive the DNet, and the goal of DNet is to separate generated image by GNet from the real image, so GNet and DNet constitute a dynamic gaming process. In the end of adversarial process, GNet can generate an image which can defeat DNet successfully, and DNet is difficult to determine whether the image is generated by GNet or not. Later Mirza and Osindero [25] adds conditions in the GAN to solve the problem of unstable training of GAN. Zhao et al. [26] proposes a generative network to resolve the problems of depth super-resolution and color super-resolution in 3D videos. Mutual information of color image and depth image are leveraged to enhance each other in consideration of the geometry structural dependency of color-depth image in the same scene. Tang and Wu [27] proposes a cascaded convolutional neural networks to implicitly learn structural information via adversarial learning for salient object detection in RGB images. Detecting the salient objects in RGB-D images using generative adversarial network is our goal. Depth information should be more considered.

3. The proposed method

The overall architecture of the cross-modal adaptive gated fusion generative adversarial network for RGB-D salient object detection is shown in Fig. 3. It consists of two parts, a generator network (GNet) with cross-modal guidance and a discriminator network (DNet) with adaptive gated fusion. GNet: first, the RGB image and depth image are fed into the double-stream network for salient object detection. Each stream takes the VGG-16 network [28] as backbone network and uses the Depthwise Separable Residual Convolution Module (DSRCM) to process deep semantic information. Then, they combine side-output features of the VGG-16 model in decoder network, and finally generate the RGB saliency map (S_{rgb}), depth saliency map (S_{depth}) and fused saliency map (S_{fusion}). DNet: first, the adaptive gated fusion module receives the last layers in the decoder of the double-stream network as the input. It guides the optimal weight fusion of RGB stream and depth stream to obtain a best gated fusion saliency map (S_{gated}). Then, the RGB image and ground-truth map (RGB+GT) are combined and sent to the discriminator network for adversarial learning. Next, RGB image and gated fusion saliency map (RGB+S_{gated}) are combined and sent to the discriminator network again. The effect of generating saliency image of GNet is improved during the iterative process, thereby obtaining a GNet that can generate clearer and higher quality saliency map. Please see the sections below for specific details of our method.

3.1. Cross-modal generator network

The generator network adopts double-stream encoder–decoder network, and the encoder network takes the VGG-16 network which pre-trained on the ImageNet [29] dataset as the backbone network. Its first five blocks are adopted and the fully connected layers are dropped for dense prediction. For decoder network, a 1 × 1 convolution is first used after the pool5 layer to reduce the feature dimension, then the reduced dimensional feature is send to the DSRCM module, which is designed based on the residual convolution module and the depthwise separable convolution module. The processed result is cascaded with the side-output features of the VGG-16 model. The next decoding operation combined with side-output features is performed in the same way. For the depth stream, it is different from the RGB stream, we use the side-output features of the RGB stream to guide the depth stream besides the same other operations. Because the feature extracted from the depth stream are not clear enough, and the reliability is not high. Therefore, the high quality feature of RGB stream is used to compensate for the drawback of depth stream. In practice, we add the side-output features of the RGB stream to the depth stream by some way, the specific way is given below. It does not lose a lot of boundaries and structural information, and enhances the feature of the depth stream. The double-stream network will eventually generate RGB saliency map, depth saliency map and fused saliency map.

3.1.1. Cross-modal guidance module

DSS [15] has achieved good performance in salient object detection for RGB images. A series of short connections from deeper to shallower side-output layers are used so that the activation of each side-output layer gains the capability of both highlighting the entire salient object and accurately depicting its boundary. Inspired by it, we use the similar approach to deal with both the RGB side-output features and the depth side-output features. But unlike DSS [15], we apply the multi-channel features instead of single-channel features to guide the side-output features of shallower layers. We also noticed that the side-output features of the depth stream are not accurate enough for the object boundary, and even interfere with the final fused saliency map, but can complement the 3D spatial information that the RGB stream cannot. Therefore, we design a simple and straightforward method to add the side-output features of the RGB stream to each side-output features of the depth stream. Short connection is also applied in the cross-modal guidance process.

Specifically, the cross-modal guidance module is designed as follows: the dimension of the side-output features of double-stream network is first reduced to 1/4 of the original feature. For the RGB stream, let us denote the side-output features of each block conv1_2, conv2_2, conv3_3, conv4_3, conv5_3, pool5 of encoder network by E^1_{rgb}, E^2_{rgb}, E^3_{rgb}, E^4_{rgb}, E^5_{rgb}, E^6_{rgb}, respectively. Similarly, the side-output features of the depth stream encoder are defined as: E^1_{depth}, E^2_{depth}, E^3_{depth}, E^4_{depth}, E^5_{depth}, E^6_{depth}. Let $\mathrm{D}^{1\sim5}_{rgb}$ denote the decoder network of the RGB stream. Mathematically, the decoder features of the RGB stream can be given by:

$$D^m_{rgb} = \begin{cases} \sum_{i=m}^{5} (E^i_{rgb}) \otimes DSRCM(D^{m+1}_{rgb}), m = 1, \ldots, 4 \\ E^m_{rgb} \otimes DSRCM(E^{m+1}_{rgb}), m = 5 \end{cases} \tag{1}$$

where DSRCM(·) represents the feature processed by the DSRCM, "⊗" denotes the operation of feature concatenation, the superscript m denotes a layer of encoder or decoder network. Similarly, let $\mathrm{D}^{1\sim5}_{depth}$ denote the decoder network of the depth stream. Because the side-output features of the RGB stream are used to guide the learning of the depth stream, so the decoder features of the depth stream can be given by:

$$D^m_{depth} = \begin{cases} \sum_{i=m}^{5} (E^i_{rgb} \otimes E^i_{depth}) \otimes DSRCM(D^{m+1}_{depth}), \\ \qquad m = 1, \ldots, 4 \\ E^m_{rgb} \otimes E^m_{depth} \otimes DSRCM(E^{m+1}_{depth}), \\ \qquad m = 5 \end{cases} \tag{2}$$

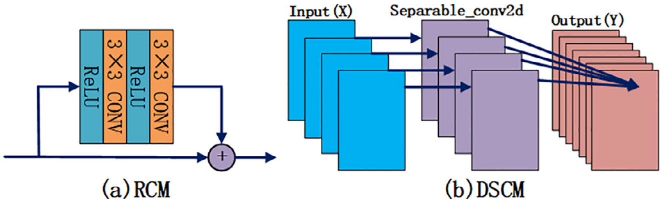


Fig. 4. The architecture of RCM and DSCM. *ReLU* and *CONV* denote the ReLU activation function and standard convolution.

In this way, the final three saliency maps of double-stream network can be expressed as:

$$S_{rgb} = sig(conv^1_{1\times1}(D^1_{rgb})) \tag{3}$$

$$S_{depth} = sig(conv^1_{1\times1}(D^1_{depth})) \tag{4}$$

$$S_{fusion} = sig(conv^1_{1\times1}(D^1_{rgb} \otimes D^1_{depth})) \tag{5}$$

where S_{rgb}, S_{depth} and S_{fusion} denote the RGB saliency map, depth saliency map and fused saliency map respectively, $conv^k_{n\times n}(\cdot)$ denotes the convolution operation using $n \times n$ convolution kernel to get k-channel features, the superscript k denotes the number of channel, $sig(\ \cdot\ )$ is the sigmoid function for generating saliency map. Three saliency maps are all supervised by ground-truth maps. The cross-modal guidance module is a reasonable design, which supplements the missing information in the depth stream, and suppresses the useless information of the depth stream to some extent.

3.1.2. Depthwise separable residual convolution module (DSRCM)

The depthwise separable residual convolution module (DSRCM) consists of two important basic components. One is the residual convolution module (RCM) [30], and the other is the depthwise separable convolution module (DSCM) [31,32]. RCM is an adaptive convolution set, as shown in Fig. 4(a). It performs the ReLU activation and 3 × 3 standard convolution of the features, and repeats this operation twice to obtain the processed features and fuse them with the original features by element addition [33]. The fusion between the residual features and the original features increases the diversity of the original features and facilitates the feature feedback process in decoder network. Therefore, RCM can enhance the feature representation from the original features. In addition, in order to reduce the parameters of the model, we introduce the depthwise separable convolution module (DSCM) which has been adopted in lightweight network ShuffeNet [34] and Xception [35] for creating an extremely efficient convolutional neural network. As shown in Fig. 4(b), the channel-level spatial convolution operation is first performed on features, and then 1 × 1 channel convolution is then performed. Its high efficiency has been verified in Table 1 when compared with standard convolution in terms of parameters and time complexity.

The depthwise separable residual convolution module (DSRCM) is designed based on RCM and DSCM, as shown in Fig. 5. First, the standard convolution module in the RCM is replaced with the DSCM to get improved RCM, then the features before and after upsampling are enhanced using the improved RCM in two different

Table 1
The difference between standard convolution and depthwise separable convolution. Suppose the following is a convolution process with an input channel of X and an output channel of Y, using $N \times N$ filter, and the size of output feature map is $M \times M$.

Convolution type	Parameter	Time complexity	Computational cost
Standard convolution	$N \times N \times X \times Y$	$O(M^2 \times N^2 \times X \times Y)$	1
Depthwise separable convolution	$(N \times N \times 1) \times X + (1 \times 1 \times X) \times Y$	$O(M^2 \times N^2 \times X + M^2 \times X \times Y)$	$1/Y + 1/N^2$

Table 2
Ablation experiments of different modules and losses.

Datasets	Evaluation Metrics	*Model –CM*	*Model –DSRCM*	*Model –GF*	*Model (DSS)*	*Model (Ls)*	*Model (Ls + La)*	*Model (Ls + Ld)*	*Model (Ls + Lg)*	*Model (Ls + La + Lg)*	*Model (Ls + Ld + Lg)*	*Model (Ls + La + Ld)*	Ours *(Ls + La + Ld + Lg)*
NLPR1000	*adpF*↑	0.8478	0.8371	0.8638	0.8345	0.8513	0.8576	0.8569	0.8593	0.8638	0.8635	0.8623	**0.8643**
	MAE ↓	0.0321	0.0411	0.0305	0.0327	0.0345	0.0332	0.0340	0.0328	0.0305	0.0317	0.0321	**0.0296**
	S↑	0.8997	0.8721	0.9043	0.8978	0.8906	0.8968	0.8934	0.8974	0.9028	0.9015	0.9005	**0.9075**
	adpE↑	0.9234	0.9117	0.9395	0.9024	0.9013	0.9029	0.9019	0.9031	0.9281	0.9271	0.9248	**0.9431**
NJU2000	*adpF*↑	0.8645	0.8369	0.8663	0.8362	0.8523	0.8582	0.8544	0.8596	0.8643	0.8639	0.8633	**0.8684**
	MAE ↓	0.0551	0.0605	**0.0515**	0.0543	0.0559	0.0549	0.0551	0.0544	0.0521	0.0529	0.0532	0.0517
	S↑	0.8786	0.8670	0.8850	0.8735	0.8721	0.8770	0.8739	0.8789	0.8843	0.8821	0.8789	**0.8851**
	adpE↑	0.9022	0.8975	0.9079	0.8966	0.8990	0.9027	0.9008	0.9030	0.9055	0.9046	0.9022	**0.9082**
STEREO	*adpF*↑	0.8574	0.8487	0.8617	0.8485	0.8564	0.8617	0.8595	0.8625	0.8624	0.8619	0.8606	**0.8638**
	MAE ↓	0.0557	0.0579	0.0512	0.0534	0.0551	0.0521	0.0534	0.0520	0.0509	0.0511	0.0518	**0.0495**
	S↑	0.8692	0.8512	0.8769	0.8702	0.8608	0.8710	0.8702	0.8721	0.8773	0.8771	0.8732	**0.8806**
	adpE↑	0.8989	0.8870	0.9021	0.8991	0.8912	0.9015	0.9004	0.9018	0.9042	0.9022	0.9011	**0.9163**
RGBD135	*adpF*↑	0.8612	0.8503	0.8702	0.8559	0.8599	0.8669	0.8630	0.8677	0.8699	0.8692	0.8671	**0.8713**
	MAE ↓	0.0357	0.0387	0.0285	0.0376	0.0351	0.0338	0.0344	0.0315	0.0301	0.0311	0.0314	**0.0282**
	S↑	0.8841	0.8713	0.9011	0.8728	0.8821	0.8891	0.8865	0.8934	0.8998	0.8984	0.8964	**0.9050**
	adpE↑	0.9075	0.8995	0.9403	0.8999	0.9022	0.9172	0.9108	0.9189	0.9338	0.9319	0.9289	**0.9421**
SIP1000	*adpF*↑	0.8126	0.7913	0.8143	0.7905	0.8005	0.8193	0.8102	0.8210	0.8274	0.8247	0.8238	**0.8294**
	MAE ↓	0.0743	0.0784	0.0723	0.0763	0.0759	0.0723	0.0736	0.0715	0.0710	0.0715	0.0719	**0.0707**
	S↑	0.8283	0.8247	0.8366	0.8211	0.8235	0.8315	0.8296	0.8321	0.8374	0.8361	0.8318	**0.8409**
	adpE↑	0.8891	0.8821	0.8906	0.8798	0.8767	0.8896	0.8826	0.8909	0.8909	0.8906	0.8901	**0.8915**
Average	*adpF*↑	0.8487	0.8329	0.8553	0.8331	0.8441	0.8527	0.8488	0.8540	0.8576	0.8566	0.8554	**0.8594**
	MAE ↓	0.0506	0.0553	0.0468	0.0509	0.0513	0.0493	0.0501	0.0484	0.0469	0.0477	0.0481	**0.0459**
	S↑	0.8720	0.8573	0.8808	0.8671	0.8658	0.8731	0.8707	0.8748	0.8803	0.8802	0.8762	**0.8838**
	adpE↑	0.9042	0.8833	0.9161	0.8956	0.8941	0.9028	0.8993	0.9035	0.9125	0.9113	0.9094	**0.9202**

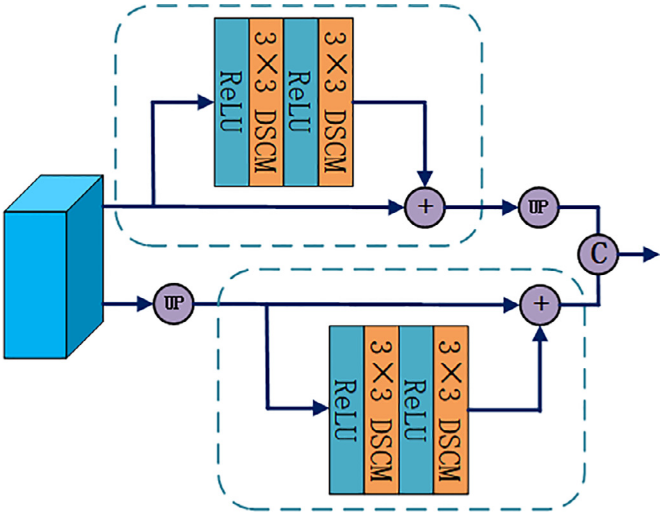


Fig. 5. The architecture of DSRCM, where *UP* denotes the operation of upsampling, and *CONV* in *RCM* is replaced with the *DSCM*.

scales. At last, two enhanced features are cascaded. On the one hand, DSRCM enhances and retains more original features by residual structure in the improved RCM. DSRCM is applied to each stage of the decoder network of double-stream network to feedback information. The high-level semantic information of the VGG-16 network is processed by DSRCM, and the information is more retained and passed to shallower layers. Moreover, enhanced features at different scales also increase the diversity of original features and improve the effect of the salient object detection. On the other hand, the standard convolution module in RCM is replaced with the DSCM, which can reduce the number of parameters in our network. Because the running time of the model is usually affected by the parameters of model, so it is also helpful to reduce the running time of our model. In summary, the design of the DSRCM plays an important role in retaining, transmitting information and reducing the number of parameters of model, so it is beneficial to improve the performance of salient object detection. The specific experimental results about DSRCM are shown in Table 2. In addition Table 4 shows the comparison of running time between our method and other methods. It embodies the advantage of DSRCM in the running time.

3.2. Adaptive gated fusion discriminator network

The adaptive gated fusion discriminator network consists of two parts, namely the adaptive gated fusion module and the discriminator module.

3.2.1. Adaptive gated fusion module

The adaptive gated fusion module is inspired by Cheng et al. [36]. Guided by the fused features of the RGB and depth streams, the gated fusion saliency map with best weight is fed into the discriminator. Specifically, the features D^1_{rgb} and D^1_{depth} from the RGB and depth streams are converted to 128-channel feature F^{128}_{rgb} and F^{128}_{depth} by convolution operation with 1 × 1 kernel. Then, they are concatenated and processed by a 1 × 1 convolution to obtain a 128-channel fused feature F^{128}_{fusion}. At last, the sigmoid layer is used to get the probability matrix $P \in \mathbb{R}^{224*224*128}$.

$$F^{128}_{rgb} = conv^{128}_{1\times 1}(D^1_{rgb}), F^{128}_{depth} = conv^{128}_{1\times 1}(D^1_{depth}) \tag{6}$$

$$F^{128}_{fusion} = conv^{128}_{1\times 1}(F^{128}_{rgb} \otimes F^{128}_{depth}) \tag{7}$$

$$P = sig(F^{128}_{fusion}) \tag{8}$$

where $conv^k_{n\times n}(\cdot)$ denotes the convolution operation using $n \times n$ convolution kernel to get k-channel features, the superscript k denotes the number of channel, "⊗" denotes the operation of feature concatenation, $sig(\cdot)$ is the sigmoid function. Let P and $1 - P$ denote the weighted gates of the RGB stream and the depth stream, respectively. So the gated fusion feature F^{128}_{gated} and the final gated fusion saliency map S_{gated} is defined as:

$$F^{128}_{gated} = P \odot F^{128}_{rgb} \oplus (1 - P) \odot F^{128}_{depth} \tag{9}$$

$$S_{gated} = sig(conv^1_{1\times 1}(F^{128}_{gated})) \tag{10}$$

Table 3
Adaptive *F*-measure, MAE, *S*-measure, adaptive *E*-measure comparisons with the state-of-the-art methods. The best three results are shown in red, blue, and green, respectively.

Datasets	Metrics	ACSD15	SE16	DF17	CTMF17	MMCI18	PCFN18	TAN19	CPFP19	Ours
NLPR1000	$adpF\uparrow$	0.5343	0.6912	0.7348	0.7234	0.7299	0.7948	**0.7956**	*0.8225*	0.8643
	$MAE\downarrow$	0.1787	0.0913	0.0891	0.0561	0.0591	0.0437	**0.0410**	*0.0359*	0.0296
	$S\uparrow$	0.6728	0.7561	0.7909	0.8599	0.8557	0.8736	**0.8861**	*0.8884*	0.9075
	$adpE\uparrow$	0.7417	0.8385	0.8600	0.8690	0.8717	**0.9163**	0.9161	*0.9240*	0.9431
NJU2000	$adpF\uparrow$	0.6964	0.7336	0.7703	0.7875	0.8122	**0.8440**	*0.8442*	0.8367	0.8684
	$MAE\downarrow$	0.2021	0.1687	0.1406	0.0847	0.0790	**0.0591**	0.0605	*0.0533*	0.0517
	$S\uparrow$	0.6992	0.6642	0.7596	0.8490	0.8581	0.8770	*0.8785*	**0.8777**	0.8851
	$adpE\uparrow$	0.7863	0.7722	0.8383	0.8638	0.8775	**0.8966**	0.8932	*0.8995*	0.9082
STEREO	$adpF\uparrow$	0.6932	0.7741	0.7650	0.7859	0.8120	**0.8450**	*0.8489*	0.8347	0.8638
	$MAE\downarrow$	0.1956	0.1452	0.1395	0.0867	0.0796	0.0606	**0.0591**	*0.0506*	0.0495
	$S\uparrow$	0.7061	0.7109	0.7664	0.8529	0.8559	*0.8800*	0.8775	**0.8798**	0.8806
	$adpE\uparrow$	0.8048	0.8308	0.8438	0.8699	0.8896	0.9054	*0.9108*	**0.9064**	0.9163
RGBD135	$adpF\uparrow$	0.7169	0.7264	0.7525	0.7777	0.7622	0.7822	**0.7948**	*0.8294*	0.8713
	$MAE\downarrow$	0.1685	0.0896	0.0933	0.0554	0.0647	0.0491	**0.0460**	*0.0379*	0.0282
	$S\uparrow$	0.7283	0.7408	0.7522	**0.8631**	0.8477	0.8418	0.8582	*0.8720*	0.9050
	$adpE\uparrow$	0.8553	0.8524	0.8775	0.9113	0.9043	0.9125	**0.9191**	*0.9273*	0.9421
SIP1000	$adpF\uparrow$	0.7270	0.6619	0.6733	0.6835	0.7946	*0.8246*	0.8087	**0.8189**	0.8294
	$MAE\downarrow$	0.1721	0.1644	0.1854	0.1394	0.0862	**0.0710**	0.0751	0.0636	*0.0707*
	$S\uparrow$	0.7316	0.6281	0.6529	0.7158	0.8329	*0.8424*	0.8347	0.8501	**0.8409**
	$adpE\uparrow$	0.8271	0.7562	0.7943	0.8239	0.8862	*0.8988*	**0.8932**	0.8990	0.8915

Table 4
Comparison of test time for each RGB-D image pair with the-state-of-the-art methods.

Method	ASCD15	SE16	DF17	CTMF17	MMCI18	PCFN18	TAN19	CPFP19	Ours
Time (s)	0.718	1.570	10.36	0.630	0.050	0.060	0.070	0.170	0.037
Code	C++	Matlab&C++	Matlab&C++	Caffe	Caffe	Caffe	Caffe	Caffe	Tensorflow

where "$\odot$" denotes Hadamard product. "$\oplus$" denotes the operation of element-wise summation. Through the adaptive gated fusion module the contribution of each modal is weighted, and better gated fusion saliency map is generated by learning the correlation of two streams.

3.2.2. Discriminator module

The discriminator module consists of three convolution blocks and three fully connected layers. The overall frame of the discriminator is shown in Fig. 3. Each convolution block consists of two convolutional layers and one max-pooling layer. The convolutional layer is activated by ReLU function, and the fully connected layer is activated by tanh function. Only the last layer uses the sigmoid activation function, and the final result is the probability whether the saliency map is true or false. If the result is greater than 0.5, it is regarded as a real image. If the result is less than 0.5, it is regarded as a false image. For the salient object detection, we use the result of the binary classification (true or false) to indicate whether the result is a ground-truth map or a saliency map produced by generator network. In the process of adversarial learning, the ground-truth map and RGB image are first fed into the discriminator as the condition of CGAN [25] to help the convergence of training. Then the gated fusion saliency map S_{gated} and RGB image are combined into the discriminator again. The discriminator is used to discriminate whether gated fusion saliency map is close to ground-truth map. The process is iteratively trained under the adversarial loss.

3.3. Loss function

The loss function consists of four parts: the adversarial loss $L_a(G, D)$, the saliency loss $L_s(G)$, the dice loss $L_d(G)$ and the gated fusion loss $L_g(G)$. It can be expressed as:

$$G^* = \min_G \max_D \alpha * L_a(G, D) + \beta * L_s(G) + \gamma * L_d(G) + \delta * L_g(G) \tag{11}$$

where the parameter α, β, γ and δ are the weight coefficients of different losses to express their importance. They are all set to 1 in our experiment. This setting of weight is to ensure that each loss contributes equally to the training of the model. As shown in Table 2, ablation experiments demonstrate the effect of various losses on the model.

3.3.1. Adversarial loss

The adversarial loss L_a is used to help the adversarial training between the discriminator network and the generator network in the GAN [24]. It can help the model to get a better generator network. X_{rgb} and X_{depth} denote RGB and depth images, Y denotes ground-truth map, Z denotes random noise vector which is implemented by dropout layer in generator network, thus the adversarial loss $L_a(G, D)$ can be expressed by:

$$\begin{aligned} L_a(G, D) = {} & \mathbb{E}_{X_{rgb}, Y} \log D(X_{rgb}, Y) \\ & + \mathbb{E}_{X_{rgb}, X_{depth}, Z} \log(1 - D(X_{rgb}, G(X_{rgb}, X_{depth}, Z))) \end{aligned} \tag{12}$$

3.3.2. Saliency loss

The saliency loss L_s is mainly defined by the cross entropy loss function. It is widely used in binary image classification and segmentation tasks. It simultaneously weights the foreground and background pixels from pixel level.

In order to fully exploit the information of different modalities, supervision on the RGB stream, depth stream and fused stream are used simultaneously. So the saliency loss consists of three parts: the saliency loss of the RGB stream, the saliency loss of the depth stream, and the saliency loss of the double-stream fusion.

$$L_s(G) = L_{rgb}(G) + L_{depth}(G) + L_{fusion}(G) \tag{13}$$

For each loss function:

$$L_i(G) = Y logS_i + (1 - Y) log(1 - S_i) \tag{14}$$

where the subscript i denotes a modality that may be rgb, depth, or fusion, S_i and Y denote the saliency map and ground-truth map respectively.

3.3.3. Dice loss

The dice loss L_d is image level loss which measures the difference between the saliency map and ground-truth map. It makes the network learn more boundary contour information. It is also used for medical image segmentation tasks, similar to IoU loss [13]. Dice loss is defined as follows:

$$L_d(G) = 1 - \frac{2\left|Y \cap S_{fusion}\right|}{|Y| + \left|S_{fusion}\right|} \tag{15}$$

where S_{fusion} and Y denote fused saliency map and ground-truth map respectively.

3.3.4. Gated fusion loss

The gated fusion loss L_g is to help the training of the gated fusion module, so that the model can learn a better gated fusion saliency map. Gated fusion loss measures the difference between the gated fusion saliency map and ground-truth map, and it is defined as follows:

$$L_g(G) = Y log S_{gated} + (1-Y) log(1-S_{gated}) \tag{16}$$

where S_{gated} and Y denote gated fusion saliency map and ground-truth map, respectively.

4. Experiment

4.1. Datasets

Five public RGB-D datasets are selected to verify the effectiveness of our method. NLPR1000 dataset [37], NJU2000 dataset [7], STEREO dataset [38], RGBD135 dataset [39] and SIP1000 dataset [40] are included.

4.1.1. NLPR1000

It consists of 1000 images with single or multiple salient objects, including common objects in various indoor and outdoor scenes under different lighting conditions.

4.1.2. NJU2000

It contains 2003 stereo image pairs and ground-truth maps with different objects, complex and challenging scenes, which are collected from 3D movies, the Internet and photos.

4.1.3. STEREO

It provides a network link for downloading stereoscopic images and manually processed ground-truth maps, including a total of 797 pairs of binocular images.

4.1.4. RGBD135

It contains 135 indoor images in different scenes, which are captured by depth cameras.

4.1.5. SIP1000

It consists of 1000 high-resolution images of multiple salient persons. The depth maps in SIP are collected by the smart phone.

4.2. Evaluation metrics

Evaluation metrics are used to measure the effectiveness of our method, including adaptive *F*-measure (adpF), mean absolute error (MAE), *S*-measure (S), adaptive E-measure (adpE) and precision–recall curve (PR curve).

4.2.1. Adaptive F-measure

Adaptive *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.

4.2.2. MAE

In order to compare the accuracy of the saliency maps, we normalize the saliency and ground-truth maps to [0,1], and calculate the mean absolute error between them, defined as the following formula:

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

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

4.2.3. S-measure

S-measure [41] simultaneously evaluates region-aware and object-aware structural similarity between a saliency map and a ground-truth map.

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

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.

4.2.4. Adaptive E-measure

Adaptive *E*-measure [42] simultaneously captures image-level statistics and local pixel matching information in enhanced-alignment matrix. It is defined as follows:

$$Q_{FM} = \frac{1}{W \times H}\sum_{x=1}^{W}\sum_{y=1}^{H}\phi_{FM}(x,y) \tag{20}$$

where ϕ_{FM} denotes the enhanced-alignment matrix described in Fan et al. [42].

4.2.5. PR curve

For a saliency map, it is first normalized using a threshold to obtain the corresponding binary mark (B); then the binary mask (B) is compared with the corresponding ground-truth (Y); finally, the average precision and recall value in the whole dataset is calculated. The PR curve is measured as follows:

$$Precision = \frac{|Y \cap B|}{|B|}, Recall = \frac{|Y \cap B|}{|Y|} \tag{21}$$

All evaluation metrics[1] are provided by the Media Computing Lab of College of Computer Science in Nankai University.

4.3. Implementation details

The resolution of input image is resized to 224 × 224. All experiments are conducted with TensorFlow [43] framework on a machine with a single NVIDIA GTX 1080Ti GPU. Adam optimizer is chosen to optimize our network parameters with a learning rate 10^{-4}. The total training time is about 43 h. No post-processing is adopted, thus the test time for each RGB-D image pair takes only 0.037 s.

Training: During the training process, for fair comparison, the same training set as Han et al. [21] is adopted, which contains 650

[1] https://mmcheng.net/zh/code-data

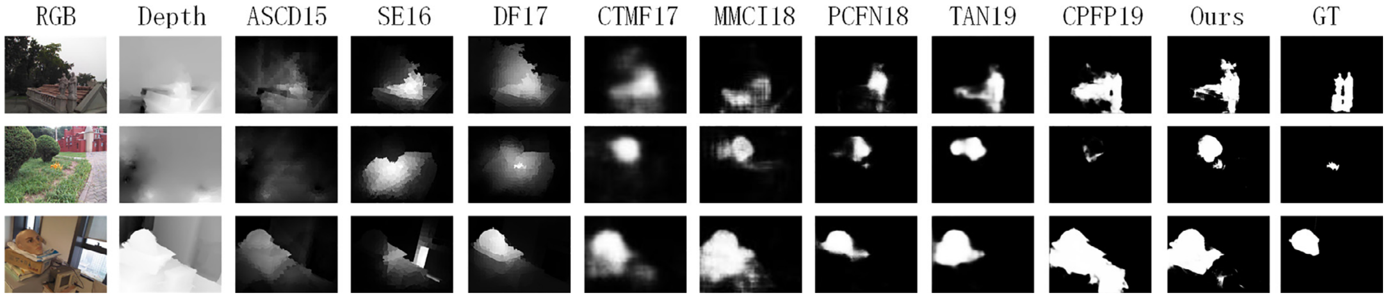


Fig. 6. Failure cases.

samples from the NLPR1000 dataset and 1400 samples from the NJU2000 dataset. It is also a training set commonly used in most current RGB-D salient object detection models. Other images in the NLPR1000 dataset and NJU2000 dataset and the entire STEREO dataset, RGBD135 dataset and SIP dataset are used for testing. Data augmentation is adopted in the training process. We simultaneously use 1 horizontally flip, 1 vertically flip, 1 translation and 1 rotation operation for RGB images, depth maps and ground-truth maps of the training set. Thus the training samples are increased by 4 times. The purpose of data augmentation is to increase the diversity of the images in training set. It is very helpful for improving the robustness of the model.

4.4. Ablation experiments

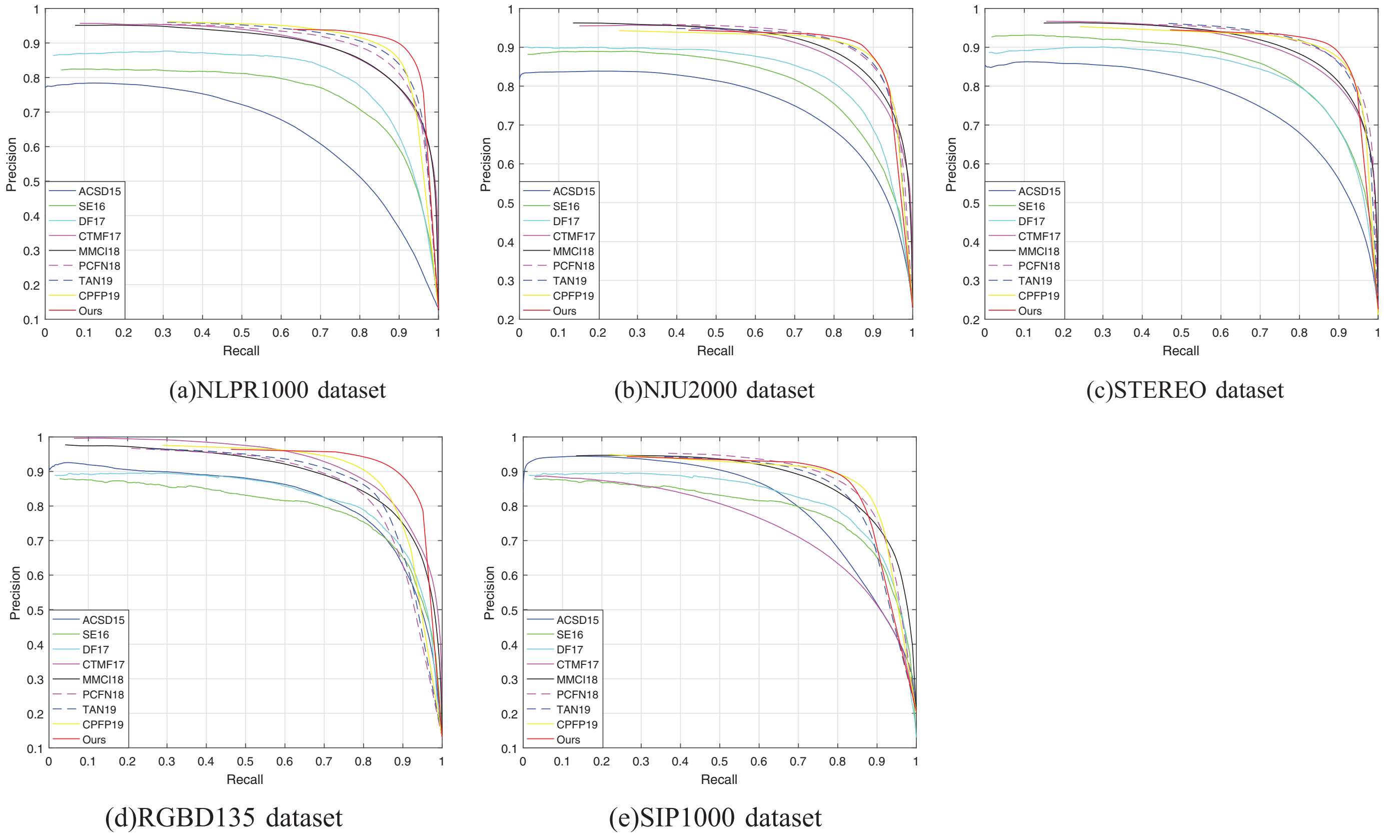
In order to verify the effectiveness of the proposed model, ablation experiments are performed based on our model denoted as *Model*. The cross-modal guidance module, the depthwise separable residual convolution module and the adaptive gated fusion module are respectively removed from the model. They are denoted as $Model - CM$, $Model - DSRCM$, $Model - GF$, respectively. For a fair comparison, the best training epoch is chosen for each model as the final result. Specifically, we choose the 136th epoch, the 127th epoch, the 114th epoch and the 117^{th} epoch as the final model of *Model*, $Model - CM$, $Model - DSRCM$, $Model - GF$, respectively. The results are shown in Table 2.

4.4.1. Cross-modal guidance module

Comparison between the third column and the last column of Table 2 shows that the *Model* which uses a cross-modal guidance in the depth stream from RGB stream is superior to $Model - CM$ which uses no cross-modal guidance. The side-output features of the RGB stream are fused to depth decoder for remedying unreliable depth feature. The guidance from RGB stream is important, and all the evaluation metrics are improved.

4.4.2. Depthwise separable residual convolution module

Comparison between the fourth column and the last column of Table 2 shows that the *Model* which uses the depthwise separable residual convolution module(DSRCM) is superior to



(a)NLPR1000 dataset

(b)NJU2000 dataset

(c)STEREO dataset

(d)RGBD135 dataset

(e)SIP1000 dataset

Fig. 7. P–R curves comparison with the state-of-the-art methods.

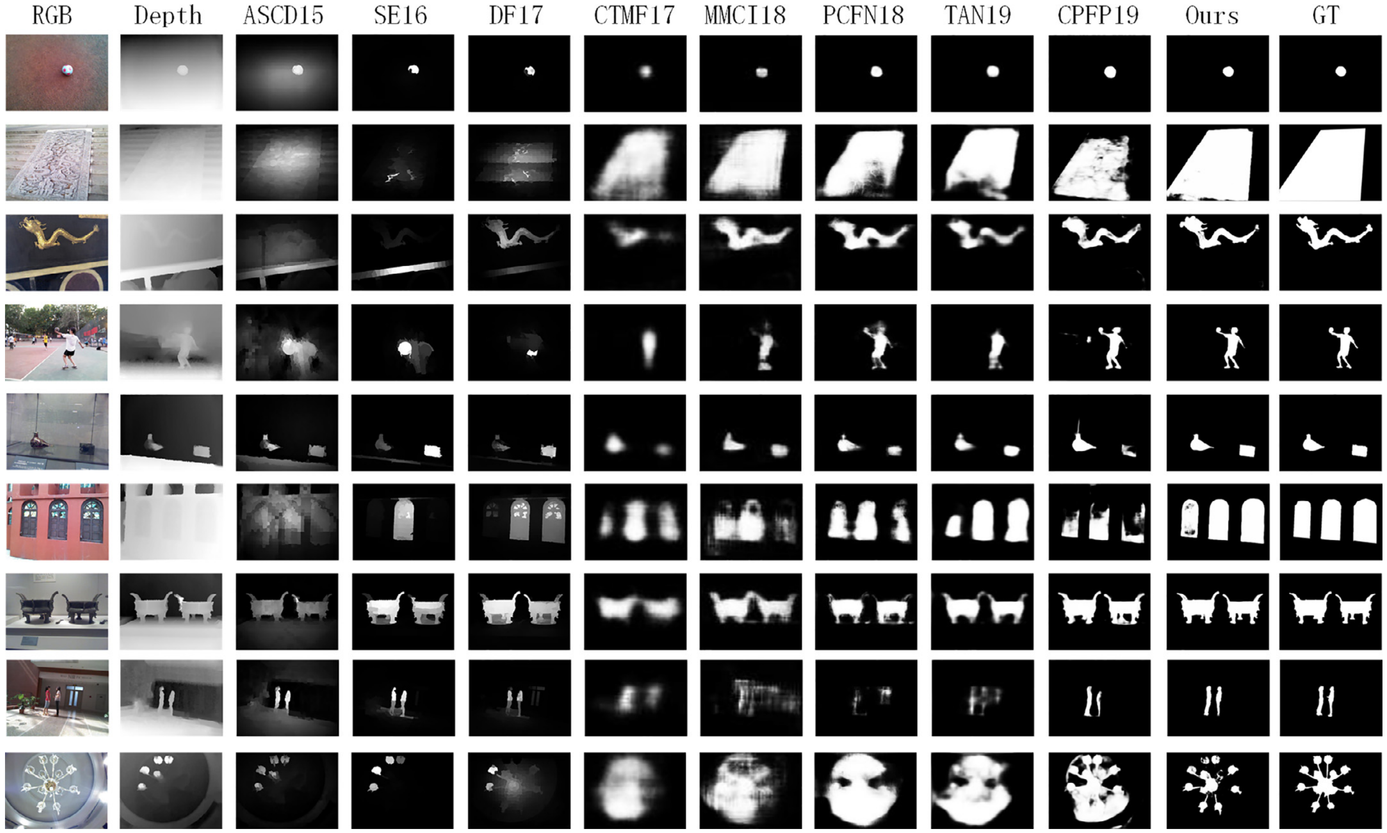


Fig. 8. Visual comparisons with the-state-of-the-art methods.

Model − DSRCM which only uses conventional upsampling operation. DSRCM is added to the decoders of double-stream network. Due to improved residual convolution module(RCM) which adopts residual structure, more abundant semantic information from deep layers is transmitted to shallower layer during the upsampling process. At the same time improved RCM is performed before and after upsampling simultaneously. They form multi-scale enhanced features which can provide the diversity of feature representation. Further the use of depthwise separable convolution module(DSCM) reduces both the parameters of the network and time cost. Table 1 shows time complexity comparison between standard convolution and depthwise separable convolution in terms of parameters, time complexity and computational cost. If the input channel of a convolution is X, its output channel is Y, the size of its filter is $N \times N$, and the size of output feature map is $M \times M$, then the number of parameters in standard convolution is $N \times N \times X \times Y$, and the number of parameters in depthwise separable convolution is $N \times N \times X + X \times Y$. The number of parameters in depthwise separable convolution is obviously less than which in standard convolution. Similarly, the time complexity is affected by the parameters. The complexity of the depthwise separable convolution is $O(M^2 \times N^2 \times X + M^2 \times X \times Y)$, which is less than the complexity of the standard convolution about $O(M^2 \times N^2 \times X \times Y)$. At last the computation cost of depthwise separable convolution is $1/Y + 1/N^2$ of standard convolutoin. So the results demonstrate that the depth separable convolution achieves the same convolution process with fewer parameters and less computational cost. In a word, comparison result proves the DSRCM plays an important role in the double-stream network.

4.4.3. Adaptive gated fusion module

Comparison between the fixth column and the last column of Table 2 shows that the *Model* which uses the adaptive gated fusion module is a little bit better than *Model − GF* which only uses element addition fusion. That is to say, the adaptive gated fusion module which is used in the discriminator network of GAN plays a minor role for our model.

Further ablation analysis on different short connection manners and different losses are experimented. The results are shown below:

4.4.4. Different short connection manner

As described in Section 3.1, the multi-channel features instead of single-channel features in DSS [15] is adopted to direct the side-output features of shallower layers. The reason is to retain more high-level semantic information and avoid the loss of information. The experimental comparison between *Model(DSS)* and *Model* in the sixth and last columns of Table 2 verifies the benefit of using multi-channel features instead of single-channel features. Our evaluation metrics are better.

4.4.5. Different losses

The different losses play different roles in our model. The saliency loss Ls is supervised on S_{rgb}, S_{depth} and S_{fusion}. It is the most basic loss of the network. The adversarial loss La is supervised on G-Net and D-Net for the training of GAN. The dice loss Ld is supervised on S_{fusion}. The gated fusion loss Lg is supervised on S_{gated}. Ablation experiments about four losses are shown in Table 2. We can see that four losses all play the important roles, and the saliency loss and gated fusion loss are more important than adversarial loss and dice loss. From the average measurement of 7th to 10th columns in the last row, we can see that the effect of different combination of losses are ranked as: Model($Ls + Lg$) $>$ Model($Ls + La$) $>$ Model($Ls + Ld$). From the average measurement of 11th to 13th columns in the last row, we can see that the effect of different combination of losses are ranked as: Model($Ls +$

$La + Lg) > \text{Model}(Ls + Ld + Lg) > \text{Model}(Ls + La + Ld)$. From 14th column, we can see that the combination of all the losses achieves the best scores.

4.5. Compare with the state-of-the-art methods

Our method is compared with several traditional salient object detection methods and deep-learning based salient object detection methods, including ACSD15 [7], SE16 [44], DF17 [45], CTMF17 [21], MMCI18 [46], PCFN18 [22], TAN19 [47], CPFP19 [23]. In order to ensure the fairness of the comparison results, the same training samples and testing samples are adopted to train and test the network model.

Evaluation metrics performance: The comparison results of all of the evaluation metrics are shown in the Table 3 and Fig. 7. Our method has better adaptive F-measure, MAE, S-measure and adaptive E-measure than the others on most datasets except for the latest public dataset SIP1000. The PR curve of our method compared with other methods has high precision and recall rate. It can be intuitively seen that our result is better than other methods.

Visual performance: The visual comparison between our method and other methods can be seen in Fig. 8. For small objects and large objects (1th and 2th row), our method has the better performance. Similarly, for the object with more background noise(3th and 4th row), our method also shows the greater advantage. For multiple salient objects (5th, 6th, 7th and 8th row) and ambiguous scenes (9th row), our method can still accurately locate salient objects and fuse the features of RGB stream and depth stream to effectively generate coherent and accurate saliency maps.

Failure cases: The proposed method has the good detection performance in most cases. But when there is a complex background in the image, or the quality of the depth map is poor, the performance of the method will be worse. Fig. 6 shows the failure cases. It is also a big challenge for current RGB-D salient object detection methods.

Comparison of running time: The proposed method has the quick detection speed about 27 fps. Due to the high efficiency of DSRCM, the test time for each RGB-D image pair takes only 0.037 seconds. Compared with the test time of other methods in Table 4, ours is the fastest on the machine with a single NVIDIA GTX 1080Ti GPU.

5. Conclusion

In this paper, we propose a cross-modal adaptive gated fusion generative adversarial network for RGB-D salient object detection. In G-Net, the feature of RGB stream guides the learning of depth stream to achieve cross-modal fusion. The depthwise separable residual convolution module transits diverse semantic feature from deep to shallow layers with fewer parameters and less time complexity. In D-Net, adaptive gated fusion module achieves adaptive fusion based on the features of double streams, and generates the better gated fusion saliency map to be delivered to discriminator. Comparisons with other existing RGB-D salient object detection methods on five publicly RGB-D datasets demonstrate that the effectiveness of the proposed method and the ability to capture salient regions in challenging situations. In the future, we will further explore RGB-D co-saliency detection which considers the inner and inter saliency constraint besides depth information.

Author Contributions

Our main contributions can be summarized as follows: (1) A cross-modal adaptive gated fusion generative adversarial network is proposed for salient object detection in RGB-D images. The generator network uses the side-output features of the RGB stream to guide the learning of the depth stream. It compensates for the shortcoming that the feature of depth map is not clear and the reliability is not high.

(2) The depthwise separable residual convolution module is proposed to upsample deep semantic information of the double-stream network. It consists of residual convolutions in which standard convolution is replaced with more efficient depthwise separable convolution. So the deep semantic information is more retained and transited to shallow layers with lower computation cost.

(3) An adaptive gated fusion is used as a part of discriminator network to adaptively fuse the features of the RGB and depth streams. It can choose the best gated fusion saliency map to input to the discriminator for adversarial learning, and further improve the effect of the generator network.

Declaration of Competing Interest

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

Acknowledgment

We would like to thank all anonymous reviewers for their valuable feedback. We further thank Prof. Ming-ming Cheng from Nankai University for providing the codes of evaluation metrics. This work is supported by National Natural Science Foundation of China (61602004), Natural Science Foundation of Anhui Province (1908085MF182) and Key Program of Natural Science Project of Educational Commission of Anhui Province (KJ2019A0034).

References

[1] 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.

[2] M.A. Turk, A.P. Pentland, Face recognition using eigenfaces, in: Proceedings of the 1991 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, IEEE, 1991, pp. 586–591.

[3] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2009) 1627–1645.

[4] D.A. Ross, J. Lim, R.-S. Lin, M.-H. Yang, Incremental learning for robust visual tracking, Int. J. Comput. Vis. 77 (1–3) (2008) 125–141.

[5] F. Radenović, G. Tolias, O. Chum, Fine-tuning CNN image retrieval with no human annotation, IEEE Trans. Pattern Anal. Mach. Intell. 41 (7) (2018) 1655–1668.

[6] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing systems, 2012, pp. 1097–1105.

[7] R. Ju, L. Ge, W. Geng, T. Ren, G. Wu, Depth saliency based on anisotropic center-surround difference, in: Proceedings of the 2014 IEEE International Conference on Image Processing (ICIP), IEEE, 2014, pp. 1115–1119.

[8] P. Huang, C.-H. Shen, H.-F. Hsiao, Rgbd salient object detection using spatially coherent deep learning framework, in: Proceedings of the 2018 IEEE Twenty-third International Conference on Digital Signal Processing (DSP), IEEE, 2018, pp. 1–5.

[9] C. Zhu, X. Cai, K. Huang, T.H. Li, G. Li, Pdnet: prior-model guided depth-enhanced network for salient object detection, arXiv:1803.08636(2018).

[10] W. Ningning, G. Xiaojin, Adaptive fusion for RGB-D salient object detection, IEEE Access 7 (2019) 55277–55284.

[11] X. Li, F. Yang, L. Chen, H. Cai, Saliency transfer: An example-based method for salient object detection., in: Proceedings of the IJCAI, 2016, pp. 3411–3417.

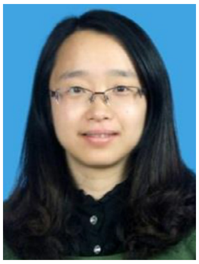
[12] G. Li, Y. Yu, Deep contrast learning for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 478–487.

[13] 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.

[14] S. Xie, Z. Tu, Holistically-nested edge detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1395–1403.

[15] 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.

[16] X. Li, F. Yang, H. Cheng, J. Chen, Y. Guo, L. Chen, Multi-scale cascade network for salient object detection, in: Proceedings of the Twenty-fifth ACM international conference on Multimedia, ACM, 2017, pp. 439–447.
[17] X. Li, F. Yang, H. Cheng, W. Liu, D. Shen, Contour knowledge transfer for salient object detection, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 355–370.
[18] L. Han, X. Li, Y. Dong, Convolutional edge constraint-based u-net for salient object detection, IEEE Access 7 (2019) 48890–48900.
[19] X. Li, K. Liu, Y. Dong, D. Tao, Patch alignment manifold matting, IEEE Trans. Neural Netw. Learn. Syst. 29 (7) (2017) 3214–3226.
[20] J. Ren, X. Gong, L. Yu, W. Zhou, M. Ying Yang, Exploiting global priors for RGB-D saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2015, pp. 25–32.
[21] 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.
[22] H. Chen, Y. Li, Progressively complementarity-aware fusion network for RGB-D salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3051–3060.
[23] 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 (CVPR), 2019, pp. 1–10.
[24] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[25] M. Mirza, S. Osindero, Conditional generative adversarial nets, arXiv:1411.1784(2014).
[26] L. Zhao, H. Bai, J. Liang, B. Zeng, A. Wang, Y. Zhao, Simultaneous color-depth super-resolution with conditional generative adversarial networks, Pattern Recognit. 88 (2019) 356–369.
[27] Y. Tang, X. Wu, Salient object detection using cascaded convolutional neural networks and adversarial learning, IEEE Trans Multimed. 21 (9) (2019) 2237–2247.
[28] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556(2014).
[29] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, L. Fei-Fei, Imagenet: A large-scale hierarchical image database, in: Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition, Ieee, 2009, pp. 248–255.
[30] S.-J. Park, K.-S. Hong, S. Lee, Rdfnet: RGB-D multi-level residual feature fusion for indoor semantic segmentation, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4980–4989.
[31] V. Vanhoucke, Learning visual representations at scale, ICLR Inv. Talk 1 (2014) 2.
[32] L. Sifre, S. Mallat, Rigid-motion scattering for image classification, Ph. D. dissertation (2014).
[33] 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.
[34] X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856.
[35] F. Chollet, Xception: Deep learning with depthwise separable convolutions, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017, pp. 1251–1258.
[36] 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.
[37] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, Rgbd salient object detection: A benchmark and algorithms, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 92–109.
[38] Y. Niu, Y. Geng, X. Li, F. Liu, Leveraging stereopsis for saliency analysis, in: Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2012, pp. 454–461.
[39] Y. Cheng, H. Fu, X. Wei, J. Xiao, X. Cao, Depth enhanced saliency detection method, in: Proceedings of the ACM ICIMCS, ACM, 2014, p. 23.
[40] D.-P. Fan, Z. Lin, J.-X. Zhao, Y. Liu, Z. Zhang, Q. Hou, M. Zhu, M.-M. Cheng, Rethinking RGB-D salient object detection: models, datasets, and large-scale benchmarks, arXiv:1907.06781(2019).
[41] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, A. Borji, Structure-measure: A new way to evaluate foreground maps, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 4548–4557.
[42] D. Fan, C. Gong, Y. Cao, B. Ren, M. Cheng, A. Borji, Enhanced-alignment measure for binary foreground map evaluation, in: International Joint Conference on Artificial Intelligence (IJCAI), 2018, pp. 698–704.
[43] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, et al., Tensorflow: Large-scale machine learning on heterogeneous distributed systems, CoRR (2016) http://arxiv.org/abs/1603.04467.
[44] J. Guo, T. Ren, J. Bei, Salient object detection for RGB-d image via saliency evolution, in: Proceedings of the 2016 IEEE International Conference on Multimedia and Expo (ICME), IEEE, 2016, pp. 1–6.
[45] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, Q. Yang, RGBD salient object detection via deep fusion, IEEE Trans. Image Process. 26 (5) (2017) 2274–2285.
[46] H. Chen, Y. Li, D. Su, Multi-modal fusion network with multi-scale multi–path and cross-modal interactions for RGB-D salient object detection, Pattern Recognit. 86 (2019) 376–385.
[47] H. Chen, Y. Li, Three-stream attention-aware network for RGB-D salient object detection, IEEE Trans. Image Process. 28 (6) (2019) 2825–2835.



Zhengyi Liu is an associate 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.

Wei Zhang is a M.S. Candidate of Anhui University. He received his B.S. from University of Science and Technology Liaoning, China in 2018. His research interests include image and video processing and computer vision.

Peng Zhao is an associate professor in School of Computer Science and Technology, Anhui University, China. She received her B.S., and M.S., from Anhui University, China in 1998 and 2003 respectively. She received Ph.D. from University of Science and Technology of China in 2006. Her research interests include image processing, and machine learning.