Salient object detection for RGB-D images by generative adversarial network

Zhengyi Liu, Jiting Tang, Qian Xiang & Peng Zhao

ONLINE FIRST

Multimedia Tools and Applications
An International Journal

ISSN 1380-7501

Multimed Tools Appl
DOI 10.1007/s11042-020-09188-8

Springer

Multimedia Tools and Applications
https://doi.org/10.1007/s11042-020-09188-8

Salient object detection for RGB-D images by generative adversarial network

Zhengyi Liu[1] · Jiting Tang[1] · Qian Xiang[1] · Peng Zhao[1]

Received: 22 July 2019 / Revised: 17 April 2020 / Accepted: 5 June 2020 /
Published online: 02 July 2020

Abstract

Salient object detection for RGB-D image aims to automatically detect the objects of human interest by color and depth information. In the paper generative adversarial network is adopted to improve its performance by adversarial learning. Generator network takes RGB-D images as inputs and outputs synthetic saliency maps. It adopts double stream network to extract color and depth feature individually and then fuses them from deep to shallow progressively. Discriminator network takes RGB image and synthetic saliency maps (RGBS), RGB image and ground truth saliency map (RGBY) as inputs, and outputs their labels indicating whether input is synthetics or ground truth. It consists of three convolution blocks and three fully connected layers. In order to pursuit long-range dependency of feature, self-attention layer is inserted in both generator and discriminator network. Supervised by real labels and ground truth saliency map, discriminator network and generator network are adversarial trained to make generator network cheat discriminator network successfully and discriminator network distinguish synthetics or ground truth correctly. Experiments demonstrate adversarial learning enhances the ability of generator network, RGBS and RGBY input in discriminator network and self-attention layer play an important role in improving the performance. Meanwhile our method outperforms state-of-the-art methods.

Keywords Generative adversarial network; Salient object detection; RGB-D image; Self-attention; Double stream network

✉ Zhengyi Liu
liuzywen@ahu.edu.cn

Jiting Tang
1796340141@qq.com

Qian Xiang
xiangqianforward@foxmail.com

Peng Zhao
18868519@qq.com

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

1 Introduction

Salient object detection (SOD) mimics human intelligence and detects the most attracting objects or instance [29, 37]. It can filter out irrelevant information and reduce the complexity of visual analysis, and it is served as an important pre-processing step in many problems, for example object detection [4, 70], VR Explorations [73] and so on. SOD research includes RGB SOD [15, 17, 83], RGB-D SOD [8, 11, 20, 39, 51, 63, 82, 85] in which depth information can be utilized, light-field SOD in which light field information is involved[38, 52, 64], high-resolution image SOD[35, 75] in which high-resolution image is directly handled, co-saliency detection [36, 68, 71, 77] in which inner and inter saliency constraint need be considered simultaneously, and video saliency detection[21, 57, 59, 65, 66] in which temporal and spatial relationship are explored. Our work focuses on RGB-D image object-aware saliency detection.

Compared with RGB image, multi-modal RGB-D image is composed of RGB color image and the paired depth image. Depth image provides more affluent spatial structure and 3D layout, which is robust against changing lighting condition. It can contribute a lot of auxiliary saliency cues especially when the salient object and background share similar appearance, or salient objects share complex color itself. Existing solutions almost adopt methods based on hand-crafted low-level features [23, 25, 32] or deep convolutional neural network (CNN) [7–9, 11, 27, 53].

Recently generative adversarial network (GAN) [24] has witnessed significant improvement in image generation [2, 3, 40],image synthesis [5, 43, 55], image inpainting [74], super-resolution [33, 49], image translation [30, 87]. GAN can learn a loss that tries to distinguish if the synthetic image is real or fake, while simultaneously training a generative model to minimize this loss. Inspired by its ability of improving the ability of generator network by adversarial learning, GAN has also been adopted in eye-fixation prediction (FP) [44, 72] or SOD for RGB images [6, 31, 45, 86]. RGBD SOD is a dense prediction task which is similar to RGB FP and RGB SOD, but it also owns some characteristics itself. How to extract the better feature from cross-modal RGB and depth image is the key problem. Meanwhile, some depth images are estimated from stereo images. They are not good. Whether the depth images with poor quality play the negative influence on the performance of GAN model is another problem. So the paper delves into these important problems and presents a RGB-D SOD model. Generator network (G-Net) first generates saliency map by double stream fusion mode in which RGB and depth image are fed into parallel feature extractors, and then merged in the highest layer, and progressively fused by color and depth feature from shallow layers and optimized by self-attention layer from deep to shallow. Saliency map generated from G-Net has achieved the better performance. Discriminator network (D-Net) receives RGB and synthetic saliency map (RGBS), RGB and ground truth saliency map (RGBY) as input, and outputs the judgement whether the input is synthetics or ground truth. Adversarial learning is applied by alternately and iteratively training D-Net and G-Net to update both weights for the best performance. When D-Net is hard to distinguish RGBS and RGBY, G-Net reaches the best state.

The contributions of our work are:

- GAN is adopted to detect salient objects in RGB-D image. G-Net receives RGB and depth image parallel and outputs synthetic saliency map. D-Net receives RGBY and RGBS and outputs binary classification indicating synthetics or ground truth. In adversarial training, G-Net generates synthetic saliency map which can easily fool D-Net, while D-Net struggles to make the right judgment. G-Net reaches the best performance when D-Net can't distinguish synthetics and ground truth correctly.

- Multi-modal fusion in GAN is probed. RGB-D image belongs to multi-modal information. Depth image is first represented as three-channel data by copying itself thrice. It is different from conventional HHA [26] representation. In G-Net, RGB and depth image are parallel processed and progressively fused by shallow layer. In D-Net, RGBY and RGBS are processed to judge whether the input is synthetics or ground truth with no depth image involved.
- Experiment results demonstrate our method outperforms state-of-the-art methods. Meanwhile, experiments also show the effect of adversarial learning, depth input with thrice original depth data in G-Net, RGBS and RGBY input in D-Net and self-attention block. Experiments further verify the selection of the shallow layers for fusion.

2 Related work

2.1 Salient object detection for RGB-D images

SOD is to detect the most attractive objects of an image. In recent years SOD for RGB image has been progressed better and better, as shown in [28, 67, 78–80, 84]. However, when the salient object and background share similar appearance, or salient objects share complex color itself, RGB image is powerless to discriminate the salient objects from background, while depth cue as complementary cue can help to detect salient objects.

This paper focuses on SOD for RGB-D images. RGB-D sensors provide an excellent ability and flexibility to capture RGB-D images in recent years. Traditional bottom-up saliency cues or hand-crafted low-level features [14, 22, 23, 25, 32, 50, 56] lack high-level semantic information and robust ability in the prevailing age of CNN. Many methods based on CNN have emerged. Chen et al. [9] first trained the saliency detection networks for RGB and depth modalities separately, and then trained a multi-modal network by fusing their deep representations in a late point. Han et al. [27] transferred the structure of the RGB-based CNN to be applicable to depth view using the task-relevant initialization and adding deep supervision in hidden layer, and fused RGB and depth views automatically. Wang et al. [69] estimated saliency from depth cues based on a CNN trained by supervision transfer, and fused the deep features of RGB images and depth images within a Bayesian framework. Chen et al. [10, 11] proposed a multi-scale multi-path fusion network with cross-modal interactions. Global and local incorporation were adopted to allow more adaptive and sufficient fusion. Chen et al. [8] proposed three-stream attention-aware multi-modal fusion network. It presented cross-modal distillation stream besides RGB-specific stream and depth-specific stream. Meanwhile, channel-wise attention was adopted in the fusion process from deep to shallow. Chen et al. [7] designed a complementarity-aware fusion (CA-Fuse) module with cross-modal residual functions and complementarity-aware supervision in the network. By cascading the CA-Fuse module and adding level-wise supervision detection network achieved the better performance. Piao et al. [51] fused multi-level RGB and depth features by residual structure. Meanwhile, depth cues with abundant spatial information was combined with multi-scale context features for accurately locating salient objects. Furthermore, a recurrent attention module was designed to boost the performance. Zhao et al. [81] enhanced depth maps based on the contrast prior, and then further combined the depth feature with RGB features by the residual connection, and last presented Fluid Pyramid Integration module to fuse the multi-scale enhanced features.

Aforementioned methods adopt CNN to achieve the better results. Inspired by the success of GAN in computer vision tasks, our work focuses more on the design of adversarial learning model for SOD in RGB-D images. G-Net can generate synthetic saliency map, and then D-Net distinguishes synthetics from ground truth. Adversarial learning can make G-Net generate the better saliency map.

2.2 Generative adversarial network for saliency detection

In recent years, GAN [24] has achieved the satisfactory performance in computer vision. The original GAN model is used to synthesize natural image with random noise vector which cannot be distinguished from real image. There is the adversarial game which can enhance the ability of G-Net for generating images closed to real images and the distinguishing ability of D-Net. Although SOD task which has the rigid ground truths that can be used to measure the performance is quite different from image generation task, the idea of adversarial learning has motivated the researches on saliency detection with GAN.

Saliency detection by GAN included FP and SOD. Yan et al. [72] proposed FP model in which feature extractor is shared by adversarial network of image generation. It mainly utilized the great ability in feature learning of GAN to extract more smooth and thorough feature, and then used transposed convolutional layers for regression to generate saliency maps. Pan et al. [44] proposed GAN to predict FP. It first trained the saliency prediction network using binary cross entropy. After this, it added the discriminator in which input was a four-channel image containing both the source image and predicted or ground truth saliency map and began adversarial training. Training proceeded alternating between training the discriminator and training the generator, by keeping the discriminator weights constant and backpropagating the error through the discriminator to update the generator's weights. At last, G-Net achieved the best performance in which output synthetic saliency map can cheat D-Net. Pan et al. [45] proposed GAN to detect salient objects in RGB image. Generator network received training images as inputs and outputed the corresponding synthetic saliency maps. Discriminator network received synthetic saliency maps or ground-truth saliency maps as inputs and outputed class labels. It first trained G-Net, and then trained D-Net, and last updated the G-Net to make it be capable to generate the better synthetic saliency maps that can deceive the D-Net. Its important contribution was conv-comparison layer which can force high-level feature of synthetic saliency maps and ground truths as similar as possible. Zhu et al. [86] presented a multi-scale adversarial feature learning model for SOD in RGB image. Its work is similar to [45]. A correlation layer was designed in the D-Net to compare the similarity between the synthetic saliency map and ground truth saliency map. Cai et al. [6] proposed the SOD model by Conditional Generative Adversarial Network (cGAN). G-Net adopted U-net structure equipped with skip connection, and D-Net was similar to DCGAN [54] dealing with original image and synthetic saliency map or ground truth saliency map pairs. Ji et al. [31] proposed the SOD model which was similar to [6], but D-Net adopted convolutional 'PatchGAN' classifier [30].

Compared with RGB FP and RGB SOD, RGBD SOD is also a dense prediction task , characteristics itself. At first, RGB and depth image are cross-modal data. How to extract the better feature from RGB stream and depth stream is the key problem. Meanwhile, some depth images are estimated from stereo images. They are not good. Whether the depth images with poor quality play the negative influence on the performance of GAN model is another problem. So the paper first designs the better model to fuse color and depth information in G-Net, and then RGB data instead of RGBD data is utilized in adversarial learning for reducing the error from depth information with poor quality.

2.3 Self-attention mechanics

Self-attention calculates the response at a position as a weighted sum of the features at all positions. It is first used in machine translation [13, 46, 61] and then applied in the vision tasks. It is complementary to convolutions and helps with modeling the long range, multi-level dependencies across the image regions. Parmar et al. [47] proposed an image transformer model to add self-attention into an autoregressive model for image generation. Wang et al. [62] formalized the self-attention as a non-local operation to model the spatial-temporal dependencies in video sequences. Zhang et al. [76] proposed the self-attention generative adversarial network (SAGAN) which learns to efficiently find global, long-range dependencies within internal representations of images. Armed with self-attention, the generator can draw images in which fine details at every location are carefully coordinated with fine details in distant portions of the image. Moreover, the discriminator can also accurately enforce complicated geometric constraints on the global image structure. Inspired by its powerful ability, self-attention block is also adopted in our G-Net and D-Net.

3 The proposed method

GAN for detecting salient objects in RGB-D images is designed and driven by generator network (G-Net) and discriminator network (D-Net), as shown in Fig. 1. G-Net takes RGB-D images as input and outputs synthetic saliency maps (S). D-Net takes RGBS (input RGB images+synthetic saliency maps) and RGBY (input RGB images+ground truth saliency maps) as inputs, and outputs their labels indicating whether input is RGBS or RGBY. Supervised by real labels and ground truth saliency maps, D-Net and G-Net are trained in an adversarial manner to make G-Net generate the better synthetic saliency maps which can cheat D-Net successfully, while D-Net tries it best to recognize synthetics or ground truth correctly. G-Net after adversarial learning can achieve the best performance, then can be used as the salient object detector for RGB-D images.

3.1 Generator network (G-Net)

G-Net adopts double stream mode to process RGB-D images. Depth image with one channel is first copied thrice to form three-channel data. RGB image and depth image are

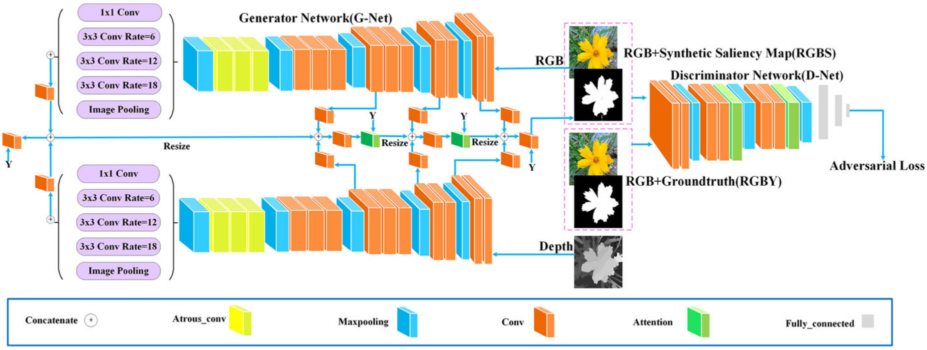


Fig. 1 Framework of GAN for SOD in RGB-D images

resized to 224×224 size and parallel fed into the feature extractor with the same structure. Each feature extractor is composed of improved VGG-16 net and improved Atrous Spatial Pyramid Pooling (ASPP) module proposed in [12]. Improved VGG-16 net replaces the 5^{th} convolutional block with an atrous convolutional block in which atrous convolutional layers have rate=2. Improved ASPP includes one 1×1 convolution, three 3×3 atrous convolutions with rates=(6,12,18) and the image-level features. It is defined as the 6^{th} block of feature extractor. It can effectively enlarge the field of view of filters for incorporating multi-scale context by four parallel atrous convolutions with different atrous rates and incorporating global context information by image-level features. The features of these five branches are then concatenated across channel direction to form global semantic feature by convolution operation. Both global semantic feature from RGB stream and depth stream are merged to form global fusion feature. The global fusion feature represents high-level semantic information of RGB-D image, and emphasizes more on the position of the salient objects, but misses the details of object outline. Thus, it needs to be fused with color and depth feature maps from shallow layers progressively to form fusion feature for sharper boundaries. Fusion with convolutional operation can adaptively combine the advantages of branches, but lack long-range dependency, so self-attention block [76] is responsible for further optimizing fusion feature. Armed with self-attention, the detail at each position of fusion feature is carefully influenced by the details in distant portions, and fusion feature is optimized by seeing more global information itself. Limited by computer memory, two self-attention blocks are adopted and inserted into the 2^{nd} and the 3^{rd} fusion parts. At last, network outputs the fusion feature optimized by color and depth features of shallow layer blocks and self-attention blocks. The deeply supervised learning scheme [34] is employed to supervise both the hidden layers and the output layer to alleviate the problem of vanishing gradients during training. Fusion features optimized in different layers are passed through another 1×1 convolution to generate the fusion saliency map. All the fusion saliency maps are supervised by ground truth saliency map, which is indicated by symbol ‘Y’ in Fig. 1. The architecture details of G-Net are presented in Table 1.

3.2 Discriminator Network (D-Net)

D-Net is used to estimate the quality of synthetic saliency map. If synthetic saliency map and ground truth saliency map are fed into D-Net, D-Net can not distinguish them, then it can be testified that synthetic saliency map generated from G-Net achieves the best performance. Therefore, D-Net should be a binary classification network which can distinguish synthetic saliency map and ground truth saliency map correctly. Meanwhile, D-Net also needs to meet the conditions that synthetic saliency map and ground truth saliency map should correspond to the same input RGB-D image.

Therefore, different from conventional GAN model [72] in which real images or fake images are fed into D-Net, and also different from models [45, 86] in which synthetic saliency map and ground truth saliency map are fed into D-Net, D-Net in our work receives input image and synthetic saliency map, input image and ground truth saliency map as input. Depth image provides auxiliary but not particularly trustworthy information [16], meanwhile, experiments 4.5.3 demonstrate that input with RGBS (RGB image and synthetic saliency map) and RGBY (RGB image and ground truth saliency map) is superior to the input with RGBDS (RGBD image and synthetic saliency map) and RGBDY (RGBD image and ground truth saliency map), so D-Net takes RGBS and RGBY as input.

Table 1 Architecture details of G-Net for detecting salient objects in RGB-D images

Block	Layer	Kernel	Padding	Strides	Output
CONV-1	2conv:conv1_1,conv1_2	3*3	Yes	1	224*224*64
	max-pool:pool1	2*2	NO	2	112*112*64
CONV-2	2conv:conv2_1,conv2_2	3*3	Yes	1	112*112*128
	max-pool:pool2	2*2	NO	2	56*56*128
CONV-3	3conv:conv3_1,conv3_2,conv3_3	3*3	Yes	1	56*56*256
	max-pool:pool3	2*2	NO	2	28*28*256
CONV-4	3conv:conv4_1,conv4_2,conv4_3	3*3	Yes	1	28*28*512
	max-pool:pool4	2*2	Yes	1	28*28*512
CONV-5	3atrous-conv:conv5_1,conv5_2,conv5_3	3*3	Yes	1	28*28*512
	max-pool:pool5	2*2	Yes	1	28*28*512
CONV-6	conv:conv6_1	1*1	Yes	1	28*28*256
	3atrous-conv:conv6_2,conv6_3,conv6_4	3*3	Yes	1	28*28*256
	global-pool+resize:pool6	2*2	Yes	1	28*28*256
CONV	concatenate+conv:conv	1*1	Yes	1	28*28*256

D-Net is composed of three convolution blocks and three fully connected layers, as shown in Fig. 1. Each convolution block includes two convolution layers and one maxpooling layer. In order to pursuit long-range dependency of feature, self-attention block [76] is inserted before maxpooling layer. Limited by computer memory, two self-attention blocks are adopted and inserted into the 2^{nd} and the 3^{rd} convolution blocks. The convolution layers employ ReLU activations, self-attention blocks adopt Leaky-ReLU activations and the fully connected layers apply tanh activations except for the final layer which uses a sigmoid activation. Architecture details of D-Net are presented in Table 2.

Table 2 Architecture details of D-Net for detecting salient objects in RGB-D images

Block	Layer	Kernel	Padding	Strides	Output
CONV-1	2conv	3*3	Yes	1	224*224*32
	max-pool	2*2	NO	2	112*112*32
CONV-2	2conv+self-attention	3*3	Yes	1	112*112*64
	max-pool	2*2	NO	2	56*56*64
CONV-3	2conv+self-attention	3*3	Yes	1	56*56*64
	max-pool	2*2	NO	2	28*28*64
fc4	fully-connected	–	–	–	100
fc5	fully-connected	–	–	–	2
fc6	fully-connected	–	–	–	1

3.3 Self-attention module

Self-attention module is used in the decoding process of our G-Net and discriminating process of our D-Net. In G-Net, feature can be refined by seeing more feature locations from global view. In D-Net, the relativity of highly detailed features in distant portions of the image is checked. The self-attention block is borrowed from reference [76]. As shown in Fig. 2, it first applies three ordinary 1 × 1 convolution operation in the input feature x and generates $k(x)$, $g(x)$ and $h(x)$ respectively. Second, the transpose output of $k(x)$ multiplies $g(x)$ and the result performs softmax normalization to form an attention map. Third, it multiplies $h(x)$ pixel by pixel to get the feature map of adaptive attention o. At last, it is multiplied by a scale parameter γ and added to the input feature map x to generate output x'.

3.4 Loss Function and Training Process

The network training includes the independent training stage of G-Net and adversarial training stage of G-Net and D-Net.

Independent Training Stage of G-Net G-Net is first trained to generate higher quality synthetic saliency maps (S). Training dataset in G-Net is denoted as $T = \{(X^n, Y^n), n = 1, \cdots, P\}$, where P denotes the number of images in the dataset. Superscript n is then dropped for notational simplicity. $X = \{x_j, j = 1, \cdots, N\}$ denotes input RGB-D image composed of RGB image $X^c = \{x_j^c, j = 1, \cdots, N\}$ and paired depth image $X^d = \{x_j^d, j = 1, \cdots, N\}$, and N denotes the number of pixels in RGB-D image X, $Y = \{y_j, j = 1, \cdots, N\}$ denotes corresponding ground truth saliency map for RGB-D image X.

Inspired by early supervision scheme [34], all the fusion saliency maps in each layer are supervised by ground truth (Y). Cross entropy loss is adopted to compare synthetic saliency map (S) with ground truth saliency map (Y) in a per-pixel manner, and defined as:

$$L_G = \sum_{m=1}^{M} [-\frac{1}{N} \sum_{j=1}^{N} (y_j log(G^{(m)}(x_j)) + (1 - y_j) log(1 - G^{(m)}(x_j)))] \tag{1}$$

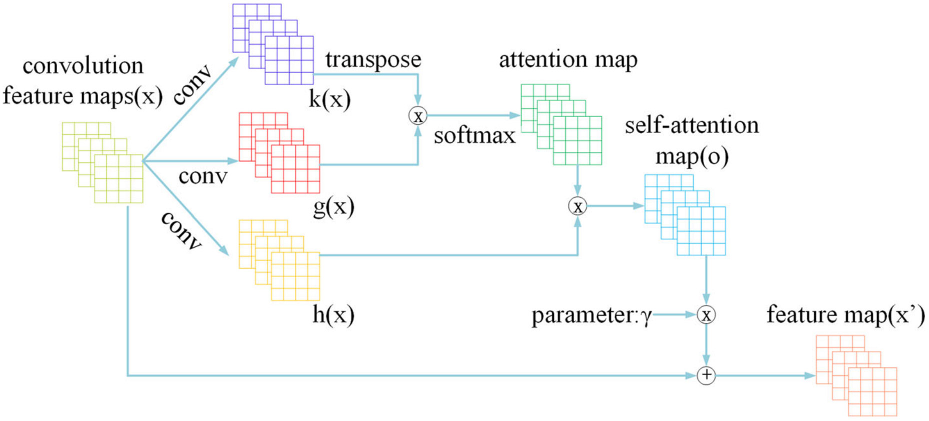


Fig. 2 Structure of self-attention block

where $G^{(m)}(x_j)$ is the probability of pixel x_j in the m^{th} fusion saliency map from G-Net being salient, M denotes the number of fusion saliency maps supervised by ground truth (Y) in G-Net, and it equals 4 in our work.

Adversarial Training Stage of G-Net and D-Net After the independent training process of G-Net, we add D-Net and begin adversarial training. SOD task belongs to image-to-image translation task where we condition on an input image and generate a corresponding output image [30]. So condition generative adversarial network (cGAN) is suitable for our task. The loss of cGAN is defined as:

$$\begin{aligned} L_{cGAN}(G, D) = \; & \mathbb{E}_{X,Y \sim P_{data}(X,Y)} \log D(X, Y) + \\ & + \mathbb{E}_{X \sim P_{data}(X), Z \sim P_Z(Z)} \log(1 - D(X, G(X, Z))) \end{aligned} \tag{2}$$

where Z denotes random noise vector which is implemented by dropout layer in G-Net. In the adversarial process G-Net tries to minimize this objective against an adversarial D-Net that tries to maximize it. Inspired by [30, 48], G-Net not only tries to fool D-Net, but also needs to be near the given ground truth in an L_1 sense to generate better saliency map, the loss is defined as:

$$L_{L_1}(G) = \mathbb{E}_{X,Y \sim P_{data}(X,Y), z \sim P_Z(Z)}[\|Y - G(X, Z)\|_1] \tag{3}$$

Therefore, the final object of GAN is:

$$G^* = \arg \min_G \max_D L_{cGAN}(G, D) + \lambda L_{L_1}(G) \tag{4}$$

In the adversarial training process, RGB-D image X is fed into G-Net, and generate synthetic saliency map $G(X, Z)$. Then RGB image X^c is combined with $G(X, Z)$ to form RGBS image $(X^c, G(X, Z))$, and RGB image X^c is combined with ground truth Y to form RGBY image (X^c, Y). Both are fed into D-Net to train D-Net by L_D. The loss of D-Net is defines as:

$$L_D = -[log D(X^c, Y) + log(1 - D(X^c, G(X, Z)))] \tag{5}$$

In order to generate a better synthetic saliency map, the weight of G-Net needs to be updated by keeping the D-Net weights constant and backpropagating the error of D-Net. So the loss of G-Net is a combination of the error from D-Net and the cross entropy loss L_G for stable and fast convergence, and defined as:

$$L = \alpha \cdot L_G + L_{GAN} \tag{6}$$

where α=0.005 is the tradeoff parameter, and L_{GAN} is the probability of fooling D-Net, and defined as:

$$L_{GAN} = -[log D(X^c, G(X, z))] \tag{7}$$

The smaller $D(X^c, G(X, z))$, the more easily that synthetic saliency map is recognized as fake by D-Net, and then the bigger L_{GAN}, the bigger the total loss L, so the weight of G-Net needs to be updating further by training.

Training in D-Net and G-Net alternately occurs. There are competitions between D-Net and G-Net. G-Net makes the network generate the better synthetic saliency map which can fool D-Net, and D-Net tries to discriminate between synthetic saliency map and ground truth saliency map. Both networks can boost their performances in the competition progressively.

4 Experimental results

4.1 Datasets

We conduct our experiments on four public benchmark datasets.

NLPR1000 [50] NLPR1000 dataset contains 1000 images captured by Microsoft Kinect. It includes 11 types of indoor and outdoor scenes and more than 400 kinds of common objects under different illumination conditions.

NJU2000 [32] NJU2000 dataset contains 2000 stereo images and photographs as well as the corresponding depth maps and manually labeled ground truths. The depth maps are generated using an optical flow method.

STERE [42] STERE dataset has 1000 stereoscopic images. These images are mainly collected from Internet and 3D movies. Depth images are estimated from disparity map of stereo images.

SIP1000 [20] SIP1000 dataset consists of 1000 high-resolution images of multiple salient persons. The depth maps in SIP are collected by the smart phone.

Training dataset adopts 1400 images from NJU2000 dataset and 650 images from NLPR1000 which is the same as CTMF [27], MMCI [11], PCFN [7] and TAN [8] for fair comparison. The remaining images are used for testing. In addition, we augment the training set by flipping all training samples horizontally.

4.2 Evaluation metrics

Evaluation metrics are used to evaluate the performance of different saliency models.

PR Curve For a saliency map, we first binarize it using a threshold to obtain the corresponding binary mark (B); then compare the binary mask (B) with the corresponding ground-truth (Y); finally, calculate the average precision and recall value in the whole dataset. The formula is as follows:

$$Precision = \frac{|Y \cap B|}{|B|}, Recall = \frac{|Y \cap B|}{|Y|} \tag{8}$$

Adaptive F-measure Adaptive F-measure is computed by:

$$F = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{9}$$

where we set $\beta^2 = 0.3$ the same as [41]. The bigger F, the better the result.

MAE Mean absolute error (MAE) refers to the average pixel-wise error between synthetic saliency map and ground truth. It is computed by:

$$MAE = \frac{\sum_{x=1}^{W}\sum_{y=1}^{H}|S(x,y) - G(x,y)|}{W \times H} \tag{10}$$

where H and W are height and width of the image, $S(x, y)$ and $G(x, y)$ denote synthetic saliency map and ground truth at the pixel (x, y) respectively. The smaller MAE, the better the result.

S-measure S-measure [18] simultaneously evaluates region-aware and object-aware structural similarity between a saliency map and a ground-truth map.

$$S_\lambda = \lambda * S_o + (1 - \lambda) * S_r \tag{11}$$

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.

Adaptive E-measure Adaptive E-measure [19] simultaneously captures image-level statistics and local pixel matching information in enhanced-alignment matrix. It is defined as follows:

$$Q_{FM} = \frac{1}{W \times H}\sum_{x=1}^{W}\sum_{y=1}^{H}\phi_{FM}(x, y) \tag{12}$$

where ϕ_{FM} denotes the enhanced-alignment matrix described in Fan et al. [19].

4.3 Implementation details

The proposed model is implemented in python with Tensorflow framework [1]. It's evaluated on a machine equipped with 1080Ti GPUs (with 32G memory). Adam optimizer is chosen to optimize our network parameters. At the independent training stage of G-Net, model is fine-tuned from an initialization with the pretrained weights of VGG-16 Network [58]. Its learning rate, weight values, bias values and mini-batch size are set as 1e-5, 0.01, 0 and 1. At the adversarial training stage of G-Net and D-Net, the learning rate is set as 1e-4, and the weight values, bias values and mini-batch size remain unchanged.

4.4 Comparison Experiment

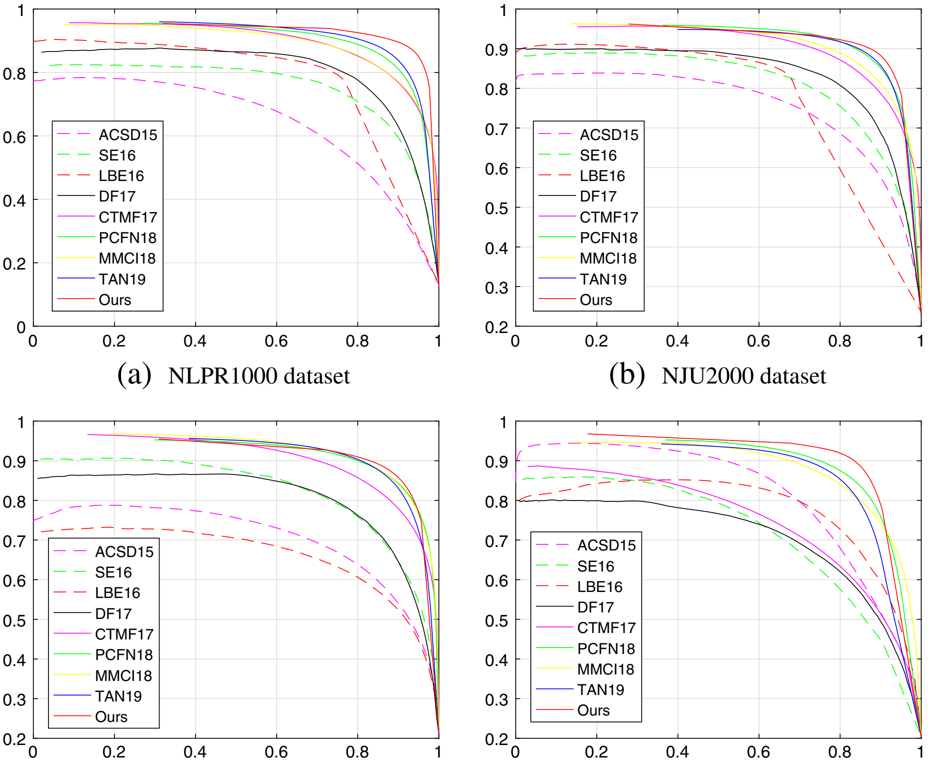
We compare our method against state-of-the-art methods ACSD15 [32], SE16 [25], LBE16 [23], DF17 [53], CTMF17 [27], PCFN18 [7], MMCI18 [11],TAN19 [8]. We use the codes provided by the authors to reproduce their results or use the result saliency maps provided by the authors. The first three methods are based on traditional hand-crafted feature extraction, and the last five methods are based on CNN.

Quantitative Evaluation Fig. 3 illustrates PR-curves of the evaluated methods. The PR curve of our method compared with other methods has high precision and recall rate on four datasets. Table 3 documents adaptive F-measure, S-measure, adaptive E-measure and MAE values. Our method consistently outperforms the state-of-the-art algorithms on four datasets in terms of all evaluation metrics, which demonstrates the effectiveness of our method.

Qualitative Evaluation Fig. 4 shows some challenging examples. Compared with the other saliency models, our method can capture entire salient objects with clear boundaries. For example, in the first and second examples, the outline of the salient objects in ours are clearer than that of the others. Meanwhile, our method works better when the foreground and background have the similar color or texture information, for example the third and fourth examples. In addition, our results are better than that of the others when not all the nearest objects to the observers are the salient objects, for example the fifth and sixth examples.

4.5 Ablation study

The proposed network is designed to achieve the better performance in SOD for RGB-D images. To show the effectiveness of each design, we take a series experiments on NLPR1000 [50], NJUD2000 [32], STERE [42] and SIP1000 [20] datasets as follows. We



(a) NLPR1000 dataset

(b) NJU2000 dataset

(c) STERE dataset

(d) SIP1000 dataset

Fig. 3 P-R curves comparison of different models

Table 3 Adaptive F-measure, MAE, S-measure and Adaptive E-measure comparisons of different models. The best three results are shown in red,blue and green,respectively

Datasets	Metrics	ACSD15	SE16	LBE16	DF17	CTMF17	PCFN18	MMCI18	TAN19	Ours
NLPR1000	$F\uparrow$	0.5343	0.6912	0.7355	0.7348	0.7234	0.7948	0.7299	0.7956	0.8572
	$MAE\downarrow$	0.1787	0.0913	0.0813	0.0891	0.0561	0.0437	0.0591	0.0410	0.0289
	$S\uparrow$	0.6728	0.7561	0.7619	0.7909	0.8599	0.8736	0.8557	0.8861	0.9103
	$E\uparrow$	0.7417	0.8385	0.8550	0.8600	0.8690	0.9163	0.8717	0.9161	0.9410
NJU2000	$F\uparrow$	0.6964	0.7336	0.7401	0.7703	0.7875	0.8440	0.8122	0.8442	0.8714
	$MAE\downarrow$	0.2021	0.1687	0.1528	0.1406	0.0847	0.0591	0.0790	0.0605	0.0521
	$S\uparrow$	0.6992	0.6642	0.6953	0.7596	0.8490	0.8770	0.8581	0.8785	0.8845
	$E\uparrow$	0.7863	0.7722	0.7913	0.8383	0.8638	0.8966	0.8775	0.8932	0.9096
STERE	$F\uparrow$	0.6607	0.7476	0.5951	0.7422	0.7708	0.8264	0.8292	0.8355	0.8528
	$MAE\downarrow$	0.2000	0.1427	0.2498	0.1409	0.0863	0.0635	0.0676	0.0596	0.0521
	$S\uparrow$	0.6919	0.7082	0.6601	0.7574	0.8480	0.8746	0.8728	0.8712	0.8763
	$E\uparrow$	0.7932	0.8250	0.7485	0.8382	0.8643	0.8967	0.9012	0.9063	0.9130
SIP1000	$F\uparrow$	0.7270	0.6619	0.7327	0.6733	0.6835	0.8246	0.7946	0.8087	0.8597
	$MAE\downarrow$	0.1721	0.1644	0.2004	0.1854	0.1394	0.0710	0.0862	0.0751	0.0608
	$S\uparrow$	0.7316	0.6281	0.7272	0.6529	0.7158	0.8424	0.8329	0.8347	0.8640
	$E\uparrow$	0.8271	0.7562	0.8405	0.7943	0.8239	0.8988	0.8862	0.8932	0.9087

take adaptive F-measure and MAE, adaptive S-measure and E-measure scores as evaluating indicators.

4.5.1 Valuation for generative adversarial network

G-Net can generate high quality saliency map itself by designed network structure, while adversarial learning can help to improve its performance further by competing with the

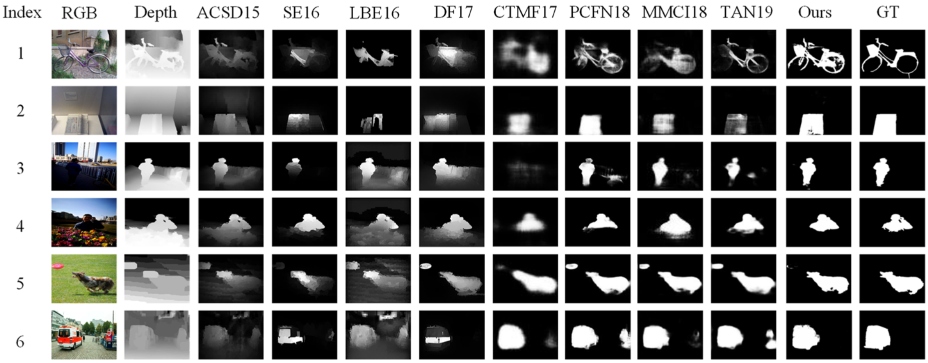


Fig. 4 Visual comparisons of different methods

D-Net. Table 4 shows that the effect of GAN. Model “G-Net” receives RGB-D image and outputs synthetic saliency map. Model “G-Net+D-Net” adds D-Net which receives RGBY and RGBS, and outputs true or false. Adversarial learning between G-Net and D-Net obviously promotes the quality of saliency maps.

4.5.2 Valuation for the input of depth image in G-Net

In the SOD methods for RGB-D image based on CNNs, depth image is almost converted to HHA representation (i.e., the horizontal disparity, height above ground, and the angle of the local surface normal with the inferred gravity direction). In our works depth image is copied thrice to form three-channel data, and then fed into depth stream network. Such selection depends on the comparison experiments shown in Table 5. We can see that all the evaluation metrics are inclined to three-channel thrice depth input.

4.5.3 Valuation for the input of D-Net

D-Net should meet the conditions that synthetic saliency map and ground truth saliency map correspond to the same input RGB-D image instead of making them indistinguishable only. Therefore, the input of D-Net should includes RGB-D image and synthetic saliency map or ground truth saliency map. But the quality of depth image in RGB-D image is sometimes unsatisfactory [60] for noise or the lost edges. So the input of D-Net is RGBS and RGBY instead of RGBDS and RGBDY in our works. In order to verify the selection, experiments about the input of D-Net are conducted, and shown in Table 6. We can see that depth can indeed have negative impact on the final performance.

4.5.4 Valuation for self-attention blocks

Self-attention blocks can model long-term dependencies in both features from G-Net and features from D-Net. Feature in G-Net can see more information in distant position by self-attention mechanics. Feature in D-Net can check the relativity of highly detailed features in distant portions of the image by self-attention mechanics. So it is inserted in G-Net and D-Net. Limited by computer memory four self-attention blocks are adopted. Two are inserted after convolutional layers in the fusion part of G-Net, and the other two are inserted in the D-Net. From Table 7 we can see performance of GAN with self-attention blocks is superior to GAN without self-attention blocks.

4.5.5 Selection for shallow layer fusion in G-Net

Global fusion feature are fused with the color and depth features from shallow layers for sharp boundaries in G-Net. If all the shallow layers are utilized for fusion, network doesn’t almost achieve the best performance. Meanwhile, as we know, shallow layer Conv1_2 from VGG-16 retains more boundary information, so it must be retained. Then experiments are conducted to decide which layers should be fused with global fusion feature as shallow layers . Table 8 shows the ablation study about the selection for fusion layers in G-Net. It is worth noting that self-attention blocks are not inserted on the consideration of speed, and adversarial learning isn’t also carried on.

From the results, we can find that selecting Conv_1,Conv_2 and Conv_3 as shallow layers to fuse with global fusion feature can achieve the best performance. At the same time, the test time for each RGB-D image pair is nearly the same about 0.021s on different groups

Table 4 Valuation for generative adversarial networks

Model	NLPR1000				NJU2000				STERE				SIP1000			
	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑
G-Net	0.8418	0.0302	0.9046	0.9297	0.8629	0.0531	0.8841	0.9044	0.8420	0.0568	0.8665	0.9104	0.8338	0.0668	0.8498	0.9041
G-Net+D-Net	**0.8572**	**0.0289**	**0.9103**	**0.9410**	**0.8714**	**0.0521**	**0.8845**	**0.9096**	**0.8528**	**0.0521**	**0.8763**	**0.9130**	**0.8597**	**0.0608**	**0.8640**	**0.9087**

The best results are in bold

Table 5 Selection for the input of depth image in G-Net

Model	NLPR1000				NJU2000				STERE				SIP1000			
	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑
HHA	0.8446	0.0308	0.9048	0.9348	0.8640	0.0534	0.8842	0.9080	0.8447	0.0546	0.8707	0.9127	0.8363	0.0645	0.8598	0.9065
Depth	**0.8572**	**0.0289**	**0.9103**	**0.9410**	**0.8714**	**0.0521**	**0.8845**	**0.9096**	**0.8528**	**0.0521**	**0.8763**	**0.9130**	**0.8597**	**0.0608**	**0.8640**	**0.9087**

The best results are in bold

Table 6 Valuation for the input of D-Net

Model	NLPR1000				NJU2000				STERE				SIP1000			
	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑
RGBD	0.8452	0.0299	0.9055	0.9370	0.8647	0.0524	0.8840	0.9072	0.8452	0.0522	0.8748	0.9112	0.8444	0.0650	0.8541	0.9021
RGB	**0.8572**	**0.0289**	**0.9103**	**0.9410**	**0.8714**	**0.0521**	**0.8845**	**0.9096**	**0.8528**	**0.0521**	**0.8763**	**0.9130**	**0.8597**	**0.0608**	**0.8640**	**0.9087**

The best results are in bold

Table 7 Valuation for self-attention blocks

Model	NLPR1000				NJU2000				STERE				SIP1000			
	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑
GAN-SA	0.8519	0.0295	0.9094	0.9404	0.8632	0.0523	0.8827	0.9046	0.8520	0.0523	0.8760	0.9004	0.8538	0.0697	0.8512	0.9032
GAN	**0.8572**	**0.0289**	**0.9103**	**0.9410**	**0.8714**	**0.0521**	**0.8845**	**0.9096**	**0.8528**	**0.0521**	**0.8763**	**0.9130**	**0.8597**	**0.0608**	**0.8640**	**0.9087**

The best results are in bold

Table 8 Selection for shallow layer fusion in G-Net

Model	NLPR1000				NJU2000				STERE				SIP1000				Test time
	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	F↑	MAE↓	S↑	E↑	
$Conv_1, Conv_2, Conv_3, Conv_4, Conv$	0.8356	0.0341	0.8954	0.9290	0.8513	0.0591	0.8731	0.9013	0.8321	0.0570	0.8601	0.9109	0.8313	0.0726	0.8442	0.8996	0.0224
$Conv_1, Conv_2, Conv$	0.8282	0.0340	0.8964	0.9233	0.8526	0.0559	0.8810	0.9004	0.8371	0.0575	0.8567	0.9072	0.8302	0.0674	0.8432	0.8985	0.0213
$Conv_1, Conv_3, Conv$	0.8415	0.0316	0.8997	0.9293	0.8553	0.0540	0.8830	0.9039	0.8410	0.0574	0.8664	0.9103	0.8309	0.0685	0.8405	0.8957	0.0206
$Conv_1, Conv_4, Conv$	0.8350	0.0342	0.8929	0.9267	0.8585	0.0545	0.8802	0.9037	0.8417	0.0571	0.8646	0.9092	0.8337	0.0672	0.8457	0.9003	**0.0200**
$Conv_1, Conv_2, Conv_4, Conv$	0.8283	0.0334	0.8963	0.9225	0.8549	0.0551	0.8803	0.9004	0.8412	0.0575	0.8660	0.9084	0.8222	0.0715	0.8456	0.8919	0.0210
$Conv_1, Conv_3, Conv_4, Conv$	0.8282	0.0348	0.8934	0.9239	0.8583	**0.0524**	0.8838	0.9011	0.8415	0.0574	0.8661	0.9088	0.8322	0.0675	0.8493	0.8999	0.0221
$Conv_1, Conv_2, Conv_3, Conv$	**0.8418**	**0.0302**	**0.9046**	**0.9297**	**0.8629**	0.0531	**0.8841**	**0.9044**	**0.8420**	**0.0568**	**0.8665**	**0.9104**	**0.8338**	**0.0668**	**0.8498**	**0.9041**	0.0210

The best results are in bold

of shallow layers. Therefore, in the situation of the same testing time, the group with best performance is selected.

5 Conclusion

In this paper, we propose a GAN for SOD in RGB-D images. Multi-modal RGB-D image is processed by double stream mode in G-Net, in which depth image is represented as three-channel thrice depth information, and fusion with feature from shallow layers is progressively performed from deep to shallow. RGBY and RGBS are processed in D-Net with no depth image involved, because depth image often contains noise. By adversarial learning of D-Net and G-Net, the distinguishable ability of D-Net and generative ability of G-Net are competed and promoted each other. The paper not only probes multi-modal fusion in GAN comprehensively, but also trains an excellent G-net to detect salient objects in RGB-D image. Experimental results demonstrate that our method can significantly outperform the state-of-the-art methods.

Acknowledgment We thank Dr. Hao Chen from City University of Hong Kong for providing their result saliency maps. We also thank Prof. Ming-ming Cheng and Dr. Deng-ping Fan from Nankai University for providing the codes of all evaluation metrics. We further thank all anonymous reviewers for their valuable comments. This research 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. Abadi M, Barham P, Chen J, Chen Z, Davis A, Dean J, Devin M, Ghemawat S, Irving G, Isard M, et al. (2016) Tensorflow: A system for large-scale machine learning. In: 12th {USENIX} symposium on operating systems design and implementation ({OSDI} 16), pp 265–283
2. Arjovsky M, Chintala S, Bottou L (2017) Wasserstein gan. arXiv:1701.07875
3. Bao J, Chen D, Wen F, Li H, Hua G (2017) Cvae-gan: fine-grained image generation through asymmetric training. In: Proceedings of the IEEE international conference on computer vision. pp 2745–2754
4. Bao J, Jia Y, Cheng Y, Xi N (2015) Saliency-guided detection of unknown objects in RGB-d indoor scenes. Sensors 15(9):21054–21074
5. Brock A, Donahue J, Simonyan K (2018) Large scale gan training for high fidelity natural image synthesis, arXiv:1809.11096
6. Cai X, Yu H (2018) Saliency detection by conditional generative adversarial network. In: Ninth international conference on graphic and image processing (ICGIP 2017), international society for optics and photonics, vol 10615, p 1061541
7. Chen H, Li Y (2018) Progressively complementarity-aware fusion network for RGB-D salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp 3051–3060
8. Chen H, Li Y (2019) Three-stream attention-aware network for RGB-d salient object detection. IEEE Trans Image Process 28(6):2825–2835
9. Chen H, Li Y, Su D (2017) RGB-D saliency detection by multi-stream late fusion network. In: International conference on computer vision systems, pp 459–468
10. Chen H, Li Y-F, Su D (2017) M3net: multi-scale multi-path multi-modal fusion network and example application to RGB-d salient object detection. In: Intelligent robots and systems (IROS). IEEE, Piscataway, pp 4911-4916
11. Chen H, Li Y, Su D (2019) Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-d salient object detection. Pattern Recogn 86:376–385
12. Chen L-C, Papandreou G, Schroff F, Adam H (2017) Rethinking atrous convolution for semantic image segmentation, arXiv:1706.05587

13. Cheng J, Dong L, Lapata M (2016) Long short-term memory-networks for machine reading, arXiv:1601.06733
14. Cheng Y, Fu H, Wei X, Xiao J, Cao X (2014) Depth enhanced saliency detection method. ACM, New York, p 23
15. Cheng M-M, Mitra NJ, Huang X, Torr PHS, Hu S-M (2015) Global contrast based salient region detection. IEEE TPAMI 37(3):569–582. https://doi.org/10.1109/TPAMI.2014.2345401
16. Cong R, Lei J, Zhang C, Huang Q, Cao X, Hou C (2016) Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. IEEE Signal Process Let, 819–823
17. Fan D-P, Cheng M-M, Liu J-J, Gao S-H, Hou Q, Borji A (2018) Salient objects in clutter: bringing salient object detection to the foreground. In: European conference on computer vision. Springer, Berlin, pp 196–212
18. Fan D-P, Cheng M-M, Liu Y, Li T, Borji A (2017) Structure-measure: A new way to evaluate foreground maps. In: Proceedings of the IEEE international conference on computer vision, pp 4558–4567
19. Fan D-P, Gong C, Cao Y, Ren B, Cheng M-M, Borji A (2018) Enhanced-alignment measure for binary foreground map evaluation, arXiv:1805.10421
20. Fan D-P, Lin Z, Zhang Z, Zhu ML, Cheng M-M (2020) Rethinking RGB-D salient object detection: models, data sets, and large-scale benchmarks. In: IEEE Transactions on neural networks and learning systems, pp 1–15
21. Fan D-P, Wang W, Cheng M-M, Shen J (2019) Shifting more attention to video salient object detection. In: IEEE CVPR, pp 8554–8564
22. Feng D, Barnes N, You S (2017) Hoso: histogram of surface orientation for RGB-d salient object detection. In: Digital image computing: techniques and applications (DICTA). IEEE, Piscataway, pp 1–8
23. Feng D, Barnes N, You S, Mccarthy C (2016) Local background enclosure for RGB-D salient object detection. In: Computer vision and pattern recognition, pp 2343–2350
24. Goodfellow I, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y (2014) Generative adversarial nets. In: Advances in neural information processing systems. pp 2672–2680
25. Guo J, Ren T, Bei J (2016) Salient object detection for RGB-D image via saliency evolution. In: IEEE International conference on multimedia and expo. pp 1–6
26. Gupta S, Girshick R, Arbeláez P, Malik J (2014) Learning rich features from RGB-D images for object detection and segmentation. In: European conference on computer vision. Springer, Berlin, pp 345–360
27. Han J, Hao C, Liu N, Yan C, Li X (2017) Cnns-based RGB-d saliency detection via cross-view transfer and multiview fusion. IEEE Trans Cybern PP(99):1–13
28. Hou Q, Cheng M-M, Hu X, Borji A, Tu Z, Torr P (2017) Deeply supervised salient object detection with short connections. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3203–3212
29. Hsu K-J, Lin Y-Y, Chuang Y-Y (2019) Deepco3: deep instance co-segmentation by co-peak search and co-saliency detection. In: IEEE CVPR, pp 8846–8855
30. Isola P, Zhu J-Y, Zhou T, Efros AA (2017) Image-to-image translation with conditional adversarial networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1125–1134
31. Ji Y, Zhang H, Wu QJ (2018) Saliency detection via conditional adversarial image-to-image network. Neurocomputing 316:357–368
32. Ju R, Ge L, Geng W, Ren T, Wu G (2014) Depth saliency based on anisotropic center-surround difference. In: Image processing ICIP. IEEE, Piscataway, pp 1115–1119
33. Ledig C, Theis L, Huszár F, Caballero J, Cunningham A, Acosta A, Aitken A, Tejani A, Totz J, Wang Z, et al. (2017) Photo-realistic single image super-resolution using a generative adversarial network. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 4681–4690
34. Lee C-Y, Xie S, Gallagher P, Zhang Z, Tu Z (2015) Deeply-supervised nets. In: Artificial intelligence and statistics, pp 562–570
35. Li C, Cong R, Hou J, Zhang S, Qian Y, Kwong S (2019) Nested network with two-stream pyramid for salient object detection in optical remote sensing images. arXiv:1906.08462
36. Li M, Dong S, Zhang K, Gao Z, Wu X, Zhang H, Yang G, Li S (2018) Deep learning intra-image and inter-images features for co-saliency detection. In: BMVC, p 291
37. Li G, Xie Y, Lin L, Yu Y (2017) Instance-level salient object segmentation. In: IEEE CVPR, pp 2386–2395
38. Li N, Ye J, Ji Y, Ling H, Yu J (2014) Saliency detection on light field. In: IEEE CVPR, pp 2806–2813
39. Liu Z, Shi S, Duan Q, Zhang W, Zhao P (2019) Salient object detection for RGB-D image by single stream recurrent convolution neural network. Neurocomputing
40. Mao X, Wang S, Zheng L, Huang Q (2018) Semantic invariant cross-domain image generation with generative adversarial networks. Neurocomputing 293:55–63

41. Martin DR, Fowlkes CC, Malik J (2004) Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Trans Pattern Anal Mach Intell 26(5):530–549
42. Niu Y, Geng Y, Li X, Liu F (2012) Leveraging stereopsis for saliency analysis. In: 2012 Computer vision and pattern recognition (CVPR) IEEE Conference on, IEEE, pp 454–461
43. Odena A, Olah C, Shlens J (2017) Conditional image synthesis with auxiliary classifier gans. In: Proceedings of the 34th international conference on machine learning-volume 70, JMLR, pp 2642–2651
44. Pan J, Canton C, Mcguinness K, O'Connor NE, Torres J, Sayrol E, Giro-i-nieto X (2017) Salgan: Visual saliency prediction with generative adversarial networks. arXiv:1701.01081
45. Pan H, Niu X, Li R, Shen S, Dou Y (2020) Supervised adversarial networks for image saliency detection. In: Eleventh international conference on graphics and image processing (ICGIP 2019), vol 11373. International Society for Optics and Photonics, p. 113730H
46. Parikh AP, Täckström O, Das D, Uszkoreit J (2016) A decomposable attention model for natural language inference, arXiv:1606.01933
47. Parmar N, Vaswani A, Uszkoreit J, Kaiser Ł, Shazeer N, Ku A, Tran D (2018) Image transformer, arXiv:1802.05751
48. Pathak D, Krahenbuhl P, Donahue J, Darrell T, Efros AA (2016) Context encoders: Feature learning by inpainting. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 2536–2544
49. Pathak HN, Li X, Minaee S, Cowan B (2018) Efficient super resolution for large-scale images using attentional gan. In: 2018 IEEE international conference on big data (Big Data). IEEE, Piscataway, pp 1777–1786
50. Peng H, Li B, Xiong W, Hu W, Ji R (2014) Rgbd salient object detection: a benchmark and algorithms. In: European conference on computer vision. Springer, Berlin, pp 92–109
51. Piao Y, Ji W, Li J, Zhang M, Lu H (2019) Depth-induced multi-scale recurrent attention network for saliency detection. In: Proceedings of the IEEE international conference on computer vision, pp 7254–7263
52. Piao Y, Rong Z, Zhang M, Li X, Lu H (2019) Deep light-field-driven saliency detection from a single view. In: IJCAI, pp 904–911
53. Qu L, He S, Zhang J, Tian J, Tang Y, Yang Q (2017) Rgbd salient object detection via deep fusion. IEEE Trans Image Process 26(5):2274–2285
54. Radford A, Metz L, Chintala S (2015) Unsupervised representation learning with deep convolutional generative adversarial networks, arXiv:1511.06434
55. Reed S, Akata Z, Yan X, Logeswaran L, Schiele B, Lee H (2016) Generative adversarial text to image synthesis, arXiv:1605.05396
56. Ren J, Gong X, Yu L, Zhou W, Ying Yang M (2015) Exploiting global priors for RGB-D saliency detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops, pp 25–32
57. Shen J, Peng J, Shao L (2018) Submodular trajectories for better motion segmentation in videos. IEEE TIP 27(6):2688–2700
58. Simonyan K, Zisserman A (2014) Very deep convolutional networks for large-scale image recognition. Computer Science
59. Song H, Wang W, Zhao S, Shen J, Lam K-M (2018) Pyramid dilated deeper convlstm for video salient object detection. In: ECCV, pp 715–731
60. Song X, Zhong F, Wang Y, Qin X (2014) Estimation of kinect depth confidence through self-training. Vis Comput 30(6-8):855–865
61. Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I (2017) Attention is all you need. In: Advances in neural information processing systems, pp 5998–6008
62. Wang X, Girshick R, Gupta A, He K (2018) Non-local neural networks. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 7794–7803
63. Wang N, Gong X (2019) Adaptive fusion for RGB-D salient object detection. arXiv:1901.01369
64. Wang T, Piao Y, Li X, Zhang L, Lu H (2019) Deep learning for light field saliency detection. In: Proceedings of the IEEE international conference on computer vision, pp 8838–8848
65. Wang W, Shen J, Shao L (2018) Video salient object detection via fully convolutional networks. IEEE TIP 27(1):38–49
66. Wang W, Shen J, Xie J, Cheng M-M, Ling H, Borji A (2019) Revisiting video saliency prediction in the deep learning era. IEEE PAMI
67. Wang L, Wang L, Lu H, Zhang P, Ruan X (2018) Salient object detection with recurrent fully convolutional networks. IEEE Trans Pattern Anal Mach Intell PP(99):1–1
68. Wang C, Zha Z-J, Liu D, Xie H (2019) Robust deep co-saliency detection with group semantic. In: AAAI, pp 8917–8924

69. Wang S, Zhou Z, Jin W, Qu H (2018) Visual saliency detection for RGB-d images under a bayesian framework. Ipsj Trans Comput Vis Appl 10(1):1
70. Wei Y (2014) Unsupervised object class discovery via saliency-guided multiple class learning. IEEE Trans Pattern Anal Mach Intell 37(4):862
71. Wei L, Zhao S, Bourahla OEF, Li X, Wu F, Zhuang Y (2019) Deep group-wise fully convolutional network for co-saliency detection with graph propagation. IEEE TIP
72. Yan B, Wang H, Wang X, Zhang Y (2017) An accurate saliency prediction method based on generative adversarial networks. In: Image processing (ICIP), 2017 IEEE international conference on, IEEE, pp 2339-2343
73. Yoon YJ, Jaechun NO, Choi SM (2017) Saliency-guided stereo camera control for comfortable vr explorations. Ieice Trans Inf Syst E100.D (9) 2245–2248
74. Yu J, Lin Z, Yang J, Shen X, Lu X, Huang TS (2018) Generative image inpainting with contextual attention
75. Zeng Y, Zhang P, Zhang J, Lin Z, Lu H (2019) Towards high-resolution salient object detection. In: IEEE ICCV, pp 1–10
76. Zhang H, Goodfellow I, Metaxas D, Odena A (2018) Self-attention generative adversarial networks, arXiv:1805.08318
77. Zhang K, Li T, Liu B, Liu Q (2019) Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing. In: IEEE CVPR, pp 3095–3104
78. Zhang P, Wang D, Lu H, Wang H, Xiang R (2017) Amulet: Aggregating multi-level convolutional features for salient object detection. In: IEEE international conference on computer vision, pp 202–211
79. Zhang X, Wang T, Qi J, Lu H, Wang G (2018) Progressive attention guided recurrent network for salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 714–722
80. Zhang P, Wang L, Wang D, Lu H, Shen C (2018) Agile amulet: real-time salient object detection with contextual attention, arXiv:1802.06960
81. Zhao J-X, Cao Y, Fan D-P, Cheng M-M, Li X-Y, Zhang L (2019) Contrast prior and fluid pyramid integration for rgbd salient object detection. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 3927–3936
82. Zhao J, Cao Y, Fan D-P, Li X-Y, Zhang L, Cheng M-M (2019) Contrast prior and fluid pyramid integration for RGBD salient object detection. In: IEEE CVPR, pp 3927–3936
83. Zhao J-X, Liu J, Fan D-P, Cao Y, Yang J, Cheng M-M (2019) EGNet: Edge guidance network for salient object detection. arXiv:1908.08297
84. Zhao R, Ouyang W, Li H, Wang X (2015) Saliency detection by multi-context deep learning. In: Proceedings of the IEEE conference on computer vision and pattern recognition, pp 1265–1274
85. Zhu C-B, Cai X, Huang K, Li TH, Li G (2019) PDNet: Prior-model guided depth-enhanced network for salient object detection. In: 2019 IEEE International conference on multimedia and expo, pp 199–204
86. Zhu D, Dai L, Luo Y, Zhang G, Shao X, Itti L, Lu J (2018) Multi-scale adversarial feature learning for saliency detection. Symmetry 10(10):457
87. Zhu J-Y, Park T, Isola P, Efros AA (2017) Unpaired image-to-image translation using cycle-consistent adversarial networks. In: Proceedings of the IEEE international conference on computer vision, pp 2223–2232

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