J. Vis. Commun. Image R. 61 (2019) 1–9

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

Depth-aware saliency detection using convolutional neural networks ☆

Yu Ding [a,b], Zhi Liu [a,b,*], Mengke Huang [a,b], Ran Shi [c], Xiangyang Wang [b]

[a] Shanghai Institute for Advanced Communication and Data Science, Shanghai University, Shanghai 200444, China
[b] School of Communication and Information Engineering, Shanghai University, Shanghai 200444, China
[c] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China

ARTICLE INFO

Article history:
Received 2 October 2018
Revised 18 January 2019
Accepted 17 March 2019
Available online 18 March 2019

Keywords:
Saliency detection
Convolutional neural networks
Depth saliency network
Saliency fusion network
RGBD images
Stereoscopic images

ABSTRACT

This paper proposes a new end-to-end depth-aware saliency model using three convolutional neural networks including color saliency network, depth saliency network and saliency fusion network, for saliency detection in RGBD images and stereoscopic images. Firstly, the color image is fed to the color saliency network to generate the color saliency map. Then, by sharing the weights of some layers in the color saliency network, the depth saliency network exploits the weight initialization and multi-layer pyramid feature fusion to learn effective depth features from the three-channel depth image, which is converted from the original depth map, and generates the depth saliency map. Finally, the saliency fusion network integrates the color saliency map with the depth saliency map into the final saliency map, which can highlight salient object regions and suppress background regions more effectively. Experimental results on five public datasets demonstrate that our model achieves the better performance compared with the state-of-the-art depth-aware saliency models.

1. Introduction

In recent years, saliency detection has become an attractive research topic in the field of computer vision. The pioneering work of Itti et al. [1] detects salient objects in images by simulating the attention mechanism of human vision. In the past decade, a large number of traditional saliency models use low-level hand-crafted features and prior knowledge, and exploit unsupervised algorithms or supervised learning to detect salient objects [2–9]. With the development of deep learning which has achieved great progresses in many fields such as image classification [10,11], as well as the augmentation of trainable labeled data, the recently proposed saliency models using convolutional neural networks have been prevalent [12–18], and a recent survey can be found in [19]. Some researchers have developed various deep neural network architectures to extract the better features for salient object detection. For example, in [16], an edge preserving and multi-scale contextual neural network produces the region-level saliency map and pixel-level saliency map for the later fusion. In [17], the proposed saliency detection framework Amulet aggregates multi-level convolutional features to generate saliency maps. In [18], the proposed end-to-end deep contrast network consists of a pixel-level fully convolutional stream and a segment-wise spatial pooling stream for saliency map generation, and a fully connected CRF model can be further incorporated to improve the spatial coherence and contour localization of saliency map. In [20], a symmetric network structure is proposed to learn complementary visual information for the better saliency detection performance. In [21], a global recursive location network is proposed to locate salient objects with more accurate context.

It should be noted that most saliency models such as those mentioned above are designed for only color images with three color channels. Although the depth information can be directly captured using depth cameras, estimated from the disparity map of stereoscopic images or even from a single image, the potential of using depth information for saliency detection has not been fully exploited. Compared with saliency detection in color images, depth-aware saliency detection for RGBD images and stereoscopic images draws less attention from the research community. Nonetheless, some depth-aware saliency models have been proposed in recent years. Based on the depth map, the anisotropic center-periphery difference [22] and the distance between the perceived location of region and the comfort zone [23] are exploited to measure saliency. In [24], region level saliency map was produced by region segmentation result, two depth features and color distribution feature. In [25], color contrast and depth contrast are enhanced with the weighting of depth-based object probability, and then region merging based saliency refinement and a location

☆ This paper has been recommended for acceptance by Dr Zicheng Liu.

* Corresponding author.

E-mail address: liuzhisjtu@163.com (Z. Liu).

https://doi.org/10.1016/j.jvcir.2019.03.019
1047-3203/© 2019 Elsevier Inc. All rights reserved.

prior of salient objects are exploited to generate the saliency map. In [26], feature contrasts of color, intensity, texture and depth are utilized for saliency detection in stereoscopic images. In [27], the Gestalt theory is introduced into saliency detection in stereoscopic videos. In [28], saliency measures were calculated by three kinds of feature such as low-level feature contrasts, mid-level feature weighted factors and high-level location priors, and then the multiscale discriminative saliency fusion is performed based on the learned random forest regressor to generate the saliency map. Li et al. [29] proposed a method that uses stereopsis to generate optical flow and proposed a new RGBD saliency detection dataset.

With the development of deep learning, a few depth-aware saliency models based on deep learning have been proposed recently. Qu et al. [30] used convolutional neural network to fuse different hand-crafted features for generating a saliency map. Han et al. [31] proposed a two-stream based framework, which uses color images and depth maps, respectively, as the input to each stream, to obtain the corresponding features, and then integrates them to generate the saliency map. Zhu et al. [32] proposed the PDNet, which is mainly based on the structure of encoder and decoder, to extract color features and depth features, fuse them and feed to the decoder to generate the saliency map. However, due to the relatively smaller amount of depth maps with the labeled ground truths, it is nontrivial to train a network for saliency estimation from depth map.

It can be seen that a common paradigm of recent depth-aware saliency models is to first extract features from color image and depth map using two networks, respectively, and then fuse the extracted features to generate the saliency map. Specifically, the fusion can take place at three different phases, *i.e.* early-phase image fusion, middle-phase feature fusion and late-phase saliency fusion. In the mode of early-phase image fusion, a color image and its depth map are directly reshaped into a 4-dimensional tensor [33], and then the tensor is fed to a neural network. Because the distributions of saliency features on depth map and color image are different, it is difficult to fit them in a network simultaneously. Therefore, the early-phase image fusion method often significantly raises the training difficulty of network. As for the mode of middle-phase feature fusion, the features of color image and depth map are respectively extracted by two different networks, and then the depth features of each layer are fused with RGB features by summation [34]. However, such a simple fusion method generates a large amount of redundant information, which degrades the saliency detection performance. In the mode of late-phase saliency fusion, the saliency map of color image and the saliency map of depth map are fused to generate the final saliency map [31,32,35]. This fusion method treats the network for color image and the network for depth map independently and only carries out information interaction at the end of the two networks by simple operations of summation or multiplication. However, such operations may not highlight salient object regions completely and suppress background regions effectively.

Motivated by the aforementioned analysis, we propose a depth-aware saliency model using convolutional neural networks to resolve the following two difficulties: (1) how to train a depth saliency network with a small amount of depth maps with the labeled ground truths, since the scale of saliency detection datasets with depth information is much smaller than the traditional saliency detection datasets for color images; (2) how to better fuse the saliency detection results from depth map and color image. The proposed depth-aware saliency model is illustrated in Fig. 1. Overall, the main contributions of our model are summarized as follows:

1) We propose a depth-aware saliency model, which consists of three networks, *i.e.* color saliency network, depth saliency network and saliency fusion network, for effective saliency detection in RGBD images and stereoscopic images.
2) The proposed depth saliency network exploits the weight initialization method to learn effective depth features with a small amount of depth data, by sharing the weights of some layers in the color saliency network to the depth saliency network.
3) The proposed saliency fusion network can effectively highlight the common salient regions in depth saliency map and color saliency map, and mutually supplement each other for the better saliency detection performance.

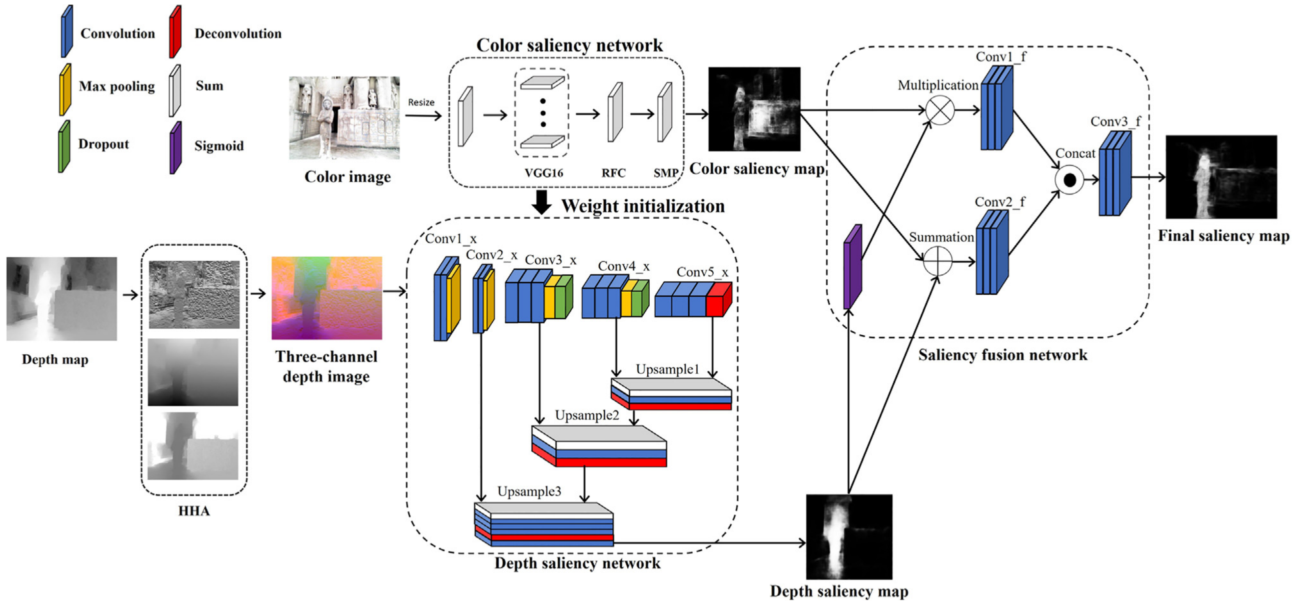


Fig. 1. Overview of the proposed depth-aware saliency model. The color saliency network takes the color image as the input and generates the color saliency map. The depth saliency network takes the three-channel depth image as the input and generates the depth saliency map. The saliency fusion network fuses the color saliency map and the depth saliency map to obtain the final saliency map.

The rest of this paper is organized as follows. Section 2 details the proposed depth-aware saliency model. Experimental results and analysis are presented in Section 3, and conclusions are drawn in Section 4.

2. Proposed depth-aware saliency model

The proposed depth-aware saliency model shown in Fig. 1 will be described in the following four subsections. Section 2.1 and Section 2.2 describe the color saliency network and the depth saliency network, respectively. Section 2.3 presents the saliency fusion network, and Section 2.4 introduces the implementation details.

2.1. Color saliency network

The color saliency network adopts an existing saliency model for color images, *i.e.* Amulet [17], and we use the parameters of its pre-trained network to initialize our color saliency network. The advantage of Amulet is that it aggregates a variety of features at different levels, which contain both semantic information and detail information. However, it is not suitable to directly use this network for extracting the depth features from the depth map, because the depth features are not as rich as the features extracted from the color image. Meanwhile, saliency detection datasets with depth information are usually smaller, and the depth maps that can be provided by these datasets for training are not enough to train the relatively larger network of Amulet. Therefore, we need to design our depth saliency network suitable for depth maps with a small amount of data for training.

2.2. Depth saliency network

For convolutional neural networks, we can use different sources of data to train a network and adjust specific structures to complete different tasks, and the difference between the parameters of the pre-trained model and the parameters of the target model is usually small in many related tasks. So, it is effective to use the color saliency network as the pre-trained model to initialize the weights of the depth saliency network for a better extraction of depth features.

Specifically, we first convert the original depth map into the three-channel depth image using the HHA algorithm [36] (HHA represents the Horizontal disparity, the Height above ground, and the Angle of the local surface normal with the inferred gravity direction). The three-channel depth image enhances complementary discontinuities in the depth map. We use the three-channel depth image as input and extract features by using VGG16 [10] as the base network, *i.e.* the layers from Conv1_x to Conv5_x. Inspired by [37], we share the weights of the VGG16 part in the color saliency network to the VGG16 part in the depth saliency network, to reasonably initialize the depth saliency network, so as to relieve the lack of training data of depth maps. For extracting the better depth features, we adjust the network architecture of VGG16 and add multi-layer feature fusion blocks to improve the feature extraction capability. Specifically, as shown in Fig. 1, we use the output feature maps of the three convolution layers in Conv2_x, Conv3_x and Conv4_x as well as the last deconvolution layer in Conv5_x, and adopt the multi-layer pyramid feature fusion architecture [38] with the operations of summation, convolution and deconvolution for the better feature fusion. To prevent network overfitting, we also add a dropout layer after the max pooling layer of Conv3_x and Conv 4_x, respectively. The specific structure of depth saliency network is shown in Table 1 and Fig. 1, in which the multi-layer pyramid feature fusion is represented using the blocks from Upsample1 to Upsample3.

Table 1
Details of depth saliency network. The convolution layers are denoted as “C[kernel size]-[number of channels]-[feature dimension]. The deconvolution layers are denoted as “DC[upsampling rate]-[number of channels]-[feature dimension].

Block	Layers
Conv1_x	C3-64-[256,256], C3-64-[256,256], Max pooling
Conv2_x	C3-128-[128,128], C3-128-[128,128], Max pooling
Conv3_x	C3-256-[64,64], C3-256-[64,64], C3-256-[64,64], Max pooling, Dropout
Conv4_x	C3-512-[32,32], C3-512-[32,32], C3-512-[32,32], Max pooling, Dropout
Conv5_x	C3-512-[16,16], C3-512-[16,16], C3-512-[16,16], DC2-512-[32,32]
Upsample1	Sum, C1-256-[32,32], DC2-256-[64,64]
Upsample2	Sum, C1-128-[64,64], DC2-128-[128,128]
Upsample3	Sum, C1-64-[128,128], C3-64-[128,128], C3-64-[128,128],DC2-64-[256,256], C1-2-[256,256]

Given the training dataset $T_{train} = \{(X_n, Y_n)\}_{n=1}^{N}$ with N training samples, in which $X_n = \{X_n^j, j = 1, \cdots, N_p\}$ and $Y_n = \{y_n^j, j = 1, \cdots, N_p\}$ denote the input image and the binary ground truth with N_p pixels, respectively. Meanwhile, $y_n^j = 1$ represent the salient object pixel and $y_n^j = 0$ represent the background pixel. Thus, the loss function for depth saliency network can be defined as:

$$L(W,b) = -\beta \sum_{j \epsilon Y_+} logP\{y^j = 1|X;W,b\} - (1-\beta) \times \sum_{j \epsilon Y_-} logP\{y^j = 0|X;W,b\} \quad (1)$$

where W and b are the kernel weights and bias of convolutional layers, and Y_+ and Y_- represent the salient objects and background label sets. The weight $\beta = |Y_+|/Y$ refers to the ratio of salient object pixels in the ground truth. $P\{y^j = 1|X;W,b\} = e^{z_1}/(e^{z_0} + e^{z_1})$ represents the probability of the pixel belonging to salient objects where z_0 and z_1 are the obtained score of background label and object label, respectively.

2.3. Saliency fusion network

In order to better fuse the color saliency map and the depth saliency map, we propose a saliency fusion network as shown in the right part of Fig. 1. First, the depth saliency map is converted to the weight map with the range of [0, 1] through the sigmoid function, which can further highlight the pixels with the higher saliency values and further suppress the pixels with the lower saliency values. The weight map is combined with the color saliency map by element-wise multiplication to obtain the saliency multiplication map, in which the common salient object regions in both color saliency map and depth saliency map will be further highlighted and the irrelevant background regions will be further suppressed. Meanwhile, the depth saliency map is also integrated with the color saliency map by element-wise summation to obtain the saliency summation map, which can recover the salient object regions only highlighted in either color saliency map or depth saliency map. After the above fusion operations of multiplication and summation, both saliency multiplication map and saliency summation map are fed into three convolution layers, and then the operations of concatenation and convolutions are performed to enable deep interactions between saliency multiplication map and saliency summation map. The details of the three convolution blocks are given in Table 2. Specifically, the final saliency map is obtained by the operations as follows:

$$S_F = Conv\{Concat[Conv(sig(S_D) \cdot S_C), Conv(S_D + S_C)]\} \quad (2)$$

Table 2
Details of the three convolution blocks in the saliency fusion network. The convolution layers are denoted as "C[kernel size]-[number of channels]-[feature dimension].

Block	Layers
Conv1_f	C3-64-[256,256], C3-64-[256,256], C1-2-[256,256]
Conv2_f	C3-64-[256,256], C3-64-[256,256], C1-2-[256,256]
Conv3_f	C3-64-[256,256], C3-64-[256,256], C1-2-[256,256]

where S_F is the final saliency map, S_D is the depth saliency map, S_C is the color saliency map, $Concat(\cdot)$ is the concatenation operation, $Conv(\cdot)$ is the convolution operation with three convolution layers, and $sig(\cdot)$ is the sigmoid function. As shown in Fig. 1, compared with both color saliency map and depth saliency map, the final saliency map generated via the saliency fusion network better highlights the salient object and suppresses background regions more effectively.

2.4. Implementation details

(1) *Data augmentation*. The scale of saliency detection datasets with depth information is smaller. Note that currently the largest public dataset NJU-2000 [22] only contains 2000 images with the labeled ground truths. Therefore, we need data augmentation to make full use of the limited data. Considering the importance of global integrity of image for saliency detection task, we use two augmentation methods. First, we flipped the images in the training set horizontally, and doubled the size of training data. Then, we rotated the original images and the flipped images by 90°, 180° and 270°, respectively, and further increased the size of training data by four times. Therefore, the training data has been expanded by eight times in total.

(2) *Loss function*. For training our saliency model, the loss function of the whole network, $Loss_A$, is defined as follows:

$$Loss_A = Loss_D + Loss_C + Loss_F \tag{3}$$

where $Loss_D$, $Loss_C$ and $Loss_F$ denote the loss of depth saliency network, color saliency network and saliency fusion network, respectively. $Loss_C$ and $Loss_F$ are defined similarly as $Loss_D$ using the same form of loss function in Eq. (1). By minimizing $Loss_A$, the saliency fusion network is fine-tuned to improve the quality of final saliency map, while the depth saliency network and color saliency network are also fine-tuned to achieve the better depth saliency map and color saliency map, which further improve the quality of saliency fusion result.

(3) *Training details*. Our saliency model was trained using the Caffe framework [39]. We trained our model with two steps. In the first step, the color saliency network is initialized by the pre-trained saliency model, *i.e.* Amulet, which also shares its weights to the depth saliency network, and all the three networks as a whole are jointly trained together by 100 iterations, to obtain the initial model with a reasonable initialization of network parameters. In the second step, all the three networks are initialized by the network parameters saved in the initial model, and further jointly trained by 14,000 iterations to obtain the final model. In details, the batch size is set to 1, the iteration size is set to 32, the learning rate is set to 10^{-8}, the dropout ratio is set to 0.5 and the momentum is set to 0.9. We implemented our saliency model on a workstation with a NVIDIA TITAN XP GPU, and it takes about 50 h to complete the whole training process.

3. Experimental results

3.1. Datasets

We evaluated the effectiveness of our depth-aware saliency model on the following five public benchmark datasets:

(1) NLPR-1000 [35]: This dataset has 1000 images captured by Microsoft Kinect stereo camera in 11 different scenes. We randomly divided the dataset into two parts: 500 images for training and 500 images for testing.
(2) NJU-2000 [22]: This dataset contains 2000 stereoscopic images from daily life, scenes and 3D movies. We randomly divided the dataset into two parts: 1500 images for training and 500 images for testing.
(3) STEREO-797 [23]: This dataset consists of 797 binocular stereo images. We used all images for testing.
(4) RGBD-135 [40]: This dataset contains 135 RGBD images from seven indoor scenarios collected by Kinect. We used all images for testing.
(5) SSD-100 [29]: This dataset consists of 80 stereo images from three stereo movies. Although this dataset is termed as SSD-100, it actually contains 80 images. We used all images for testing.

3.2. Evaluation metrics

We use the most widely used three evaluation metrics, *i.e.* precision-recall (PR) curve, F-measure and mean absolute error (MAE) to evaluate the saliency detection performance of different models. We binarize the saliency map using each integer threshold from 0 to 255. By comparing the binary saliency map with the corresponding ground truth, we can calculate the precision and recall at each integer threshold to plot the PR curve. F-measure is an overall metric to quantitatively measure saliency detection performance. To calculate F-measure, the binary object mask is obtained by adaptively thresholding the saliency map using [41], the precision and recall are then calculated by comparing the binary object mask with the ground truth, and the F-measure is finally calculated as follows:

$$F_\beta = \frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times precision \times recall}, \tag{4}$$

where $\beta^2 = 1$ puts the same weight to precision and recall. Moreover, we also use MAE to test the robustness of saliency model. The MAE computes the difference at pixel level between the saliency map S and the ground truth G, and is defined as follows:

$$MAE = \frac{1}{W \times H} \sum\nolimits_{x=1}^{W} \sum\nolimits_{y=1}^{H} |S(x,y) - G(x,y)| \tag{5}$$

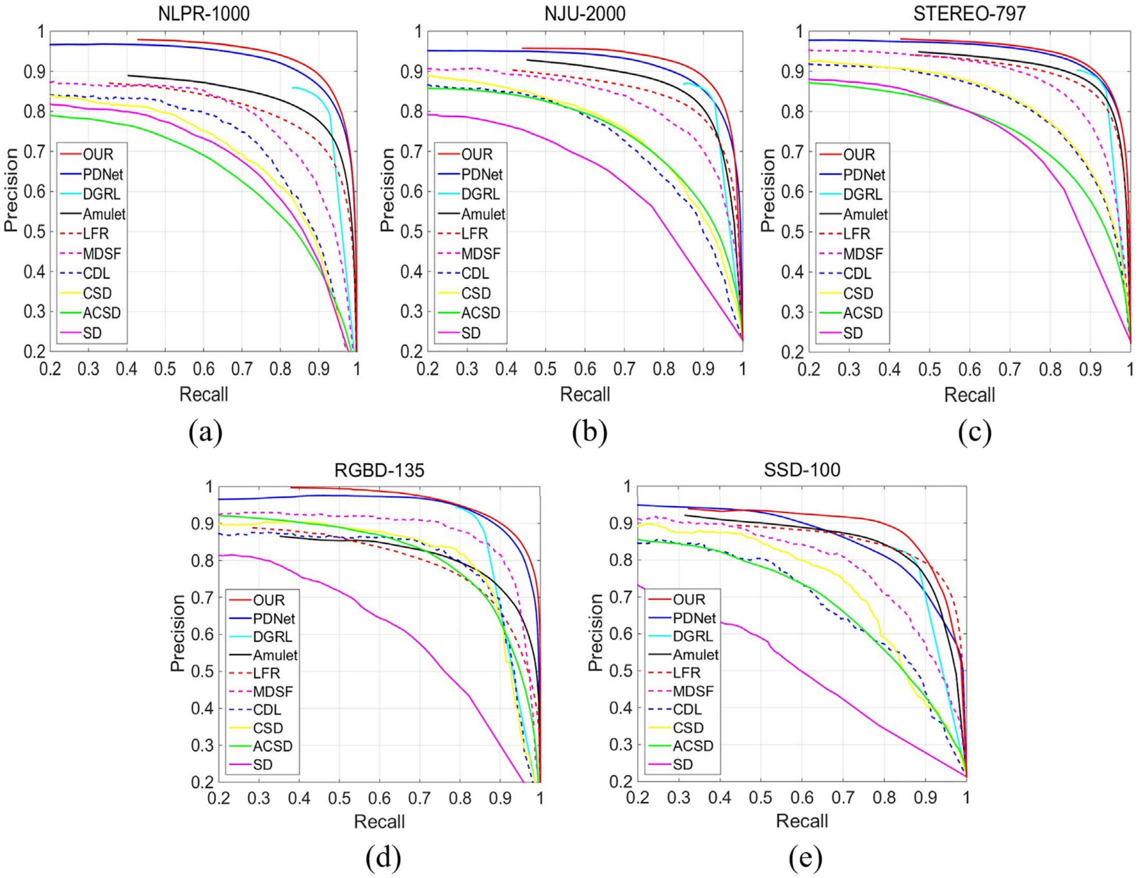
where W and H are the width and height, respectively, of the saliency map S. A smaller value of MAE indicates a better saliency detection performance.

3.3. Quantitative comparison

We compared the saliency detection performance with six state-of-the-art depth-aware saliency models including PDNet [32], SD [35], ACSD [22], CSD [24], CDL [25] and MDSF [28] as well as three state-of-the-art image saliency models including Amulet [17], LFR [20] and DGRL [21] For all the above nine saliency models, we used the source codes or the experimental results provided by their authors. We generated the saliency maps for all test images using the proposed depth-aware saliency model and the above nine saliency models. The PR curves plotted for all saliency models are shown in Fig. 2. The average F-measure and average MAE achieved by all saliency models are shown in Table 3. We can see from Fig. 2 and Table 3 that our saliency model outperforms all the other saliency models on all the five datasets in terms of all the three evaluation metrics.

3.4. Qualitative comparison

Some saliency maps are shown in Figs. 3–6 for a qualitative comparison. Overall, the saliency maps generated using our model show the best visual quality compared with those generated using all the other saliency models. It can be seen from Figs. 3–6 that our model suppresses background regions more effectively and highlights salient objects more uniformly with well-defined boundaries. For example, in the 1st row of Fig. 3, the shape of salient object is complicated. By extracting the detailed edge features of salient object from the depth map, our model can generally better highlight the salient object with well-defined boundaries. When the object color is not prominent or the background is with the complicated scene, such as the two examples in Fig. 4, our saliency model can still delineate the most accurate salient objects. For images with the cluttered background, such as the 2nd row of Fig. 3 and the 2nd row of Fig. 6, some background regions are falsely highlighted by other saliency models, while our model can better suppress such cluttered background regions via the effective utilization of depth maps.



(a) (b) (c)

(d) (e)

Fig. 2. Comparison of precision-recall (PR) curves among different saliency models on the five datasets.

Table 3
Comparison of average F-measure and average MAE among different saliency models on the five datasets.

Dataset	Metric	OUR	PDNet	DGRL	Amulet	LFR	MDSF	CDL	CSD	ACSD	SD
NLPR-1000	F-measure	**0.882**	0.865	0.837	0.793	0.757	0.638	0.580	0.575	0.594	0.542
	MAE	**0.038**	0.050	0.043	0.065	0.088	0.124	0.117	0.276	0.167	0.111
NJU-2000	F-measure	**0.884**	0.864	0.859	0.827	0.810	0.741	0.653	0.685	0.683	0.375
	MAE	**0.054**	0.071	0.058	0.082	0.093	0.150	0.174	0.281	0.189	0.200
STEREO-797	F-measure	**0.896**	0.889	0.886	0.873	0.856	0.822	0.748	0.731	0.706	0.473
	MAE	0.049	0.062	**0.044**	0.060	0.069	0.128	0.210	0.285	0.184	0.183
RGBD-135	F-measure	**0.894**	0.880	0.840	0.777	0.742	0.748	0.712	0.731	0.701	0.418
	MAE	**0.029**	0.039	0.031	0.068	0.088	0.097	0.086	0.252	0.153	0.114
SSD-100	F-measure	**0.846**	0.771	0.812	0.805	0.810	0.705	0.603	0.651	0.617	0.405
	MAE	**0.059**	0.107	0.061	0.085	0.083	0.163	0.173	0.276	0.201	0.205

The bold font indicates the best result in this row.

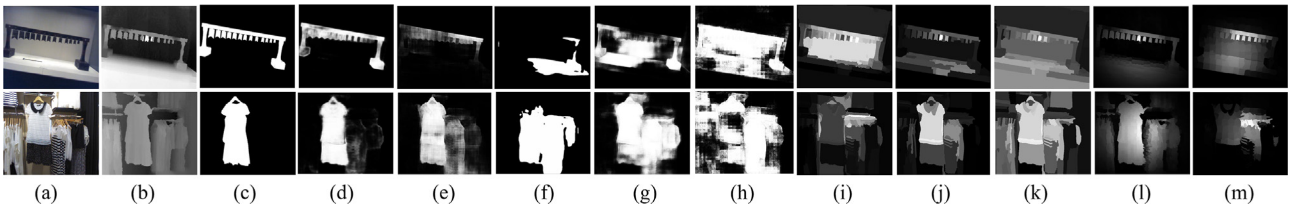


Fig. 3. Visual comparison of saliency maps on the NLPR-1000 dataset. (a) color image; (b) depth map; (c) ground truth; saliency maps generated using (d) our model, (e) PDNet, (f) DGRL, (g) Amulet, (h) LFR, (i) MDSF, (j) CDL, (k) CSD, (l) ACSD, (m) SD, respectively.

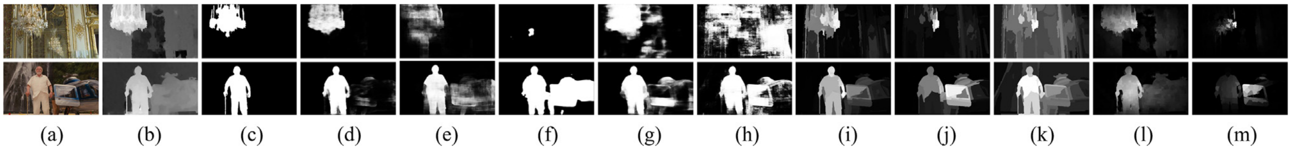


Fig. 4. Visual comparison of saliency maps on the NJU-2000 dataset. (a) color image; (b) depth map; (c) ground truth; saliency maps generated using (d) our model, (e) PDNet, (f) DGRL, (g) Amulet, (h) LFR, (i) MDSF, (j) CDL, (k) CSD, (l) ACSD, (m) SD, respectively.

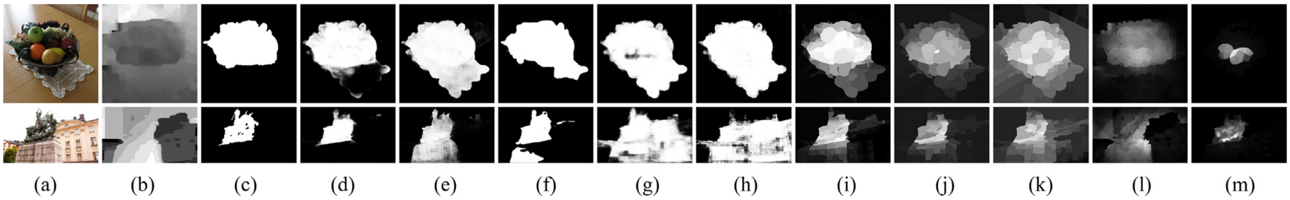


Fig. 5. Visual comparison of saliency maps on the STEREO-797 dataset. (a) color image; (b) depth map; (c) ground truth; saliency maps generated using (d) our model, (e) PDNet, (f) DGRL, (g) Amulet, (h) LFR, (i) MDSF, (j) CDL, (k) CSD, (l) ACSD, (m) SD, respectively.

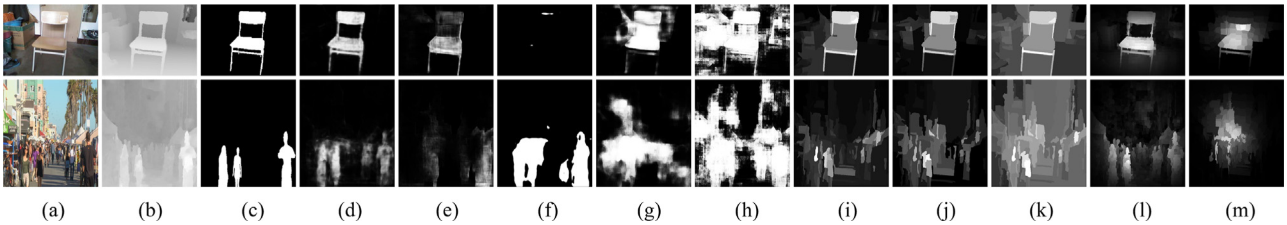


Fig. 6. Visual comparison of saliency maps on the RGBD-135 and SSD-100 dataset. (a) color image; (b) depth map; (c) ground truth; saliency maps generated using (d) our model, (e) PDNet, (f) DGRL, (g) Amulet, (h) LFR, (i) MDSF, (j) CDL, (k) CSD, (l) ACSD, (m) SD, respectively.

3.5. Ablation study

In this subsection, we will evaluate the effectiveness of our model through comprehensive ablation experiments. We performed ablation experiments on all the five datasets. For conciseness, here we only show the results on the NLPR-1000 dataset. We also observed the three evaluation metrics on the other four datasets and can draw the same conclusions as those in the following paragraphs.

1) To analyze the contributions of depth saliency network, saliency fusion network and weight initialization method, as shown in Table 4 and Fig. 8(a), we compared our final saliency maps with three classes of saliency maps, *i.e.* final saliency maps generated using our model without weight initialization, color saliency maps and depth saliency maps. They are denoted as "OUR", "WOWI", "Color" and "Depth", respectively, in Table 4 and Fig. 8(a). Here, the color saliency maps and depth saliency maps are generated by using the independently trained color saliency network (Section 2.1) and depth saliency network (Section 2.2), respectively. It can be seen that in terms of all the three evaluation metrics, our final saliency maps are consistently better than the other three classes of saliency maps. This demonstrates the effec-

Table 4
Comparison of average F-measure and average MAE among different model settings and fusion methods on the NLPR-1000 dataset.

Metric	F-measure	MAE
OUR	**0.882**	**0.038**
WIWO	0.871	0.042
Color	0.834	0.047
Depth	0.731	0.094
OUR-LFR	**0.858**	**0.045**
LFR	0.757	0.088
OUR	**0.882**	**0.038**
OUR-DM	0.860	0.044
OUR	**0.882**	**0.038**
PNSP	0.850	0.064
Sum	0.849	0.065
Prod	0.820	0.059
Max	0.819	0.068

tiveness of the proposed depth saliency network, saliency fusion network and the weight initialization method to progressively improve the saliency detection performance.

Besides, an example is shown in Fig. 7 to illustrate the contributions of depth saliency network, saliency fusion network and the weight initialization method in our model. For the example image in Fig. 7(a), in which the color contrast between the salient object and background regions is very low, the color saliency map generated by using the independently trained color saliency network is shown in Fig. 7(b), which cannot distinguish the salient object from background. The depth saliency map generated by using the independently trained depth saliency network is shown in Fig. 7(c), which can mostly highlight the salient object regions, but without accurate object boundaries. As shown in Fig. 7(d), without the weight initialization, the background regions cannot be effectively suppressed in the final saliency map. For a visual comparison, as shown in Fig. 7(e), the salient object with well-defined boundaries

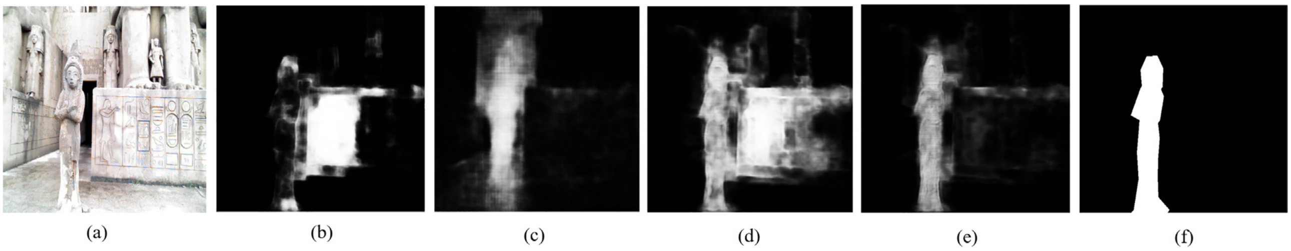


Fig. 7. Results of ablation study. (a) color image; (b) color saliency map; (c) depth saliency map; (d) final saliency map without weight initialization; (e) final saliency map generated using our model; (f) ground truth.

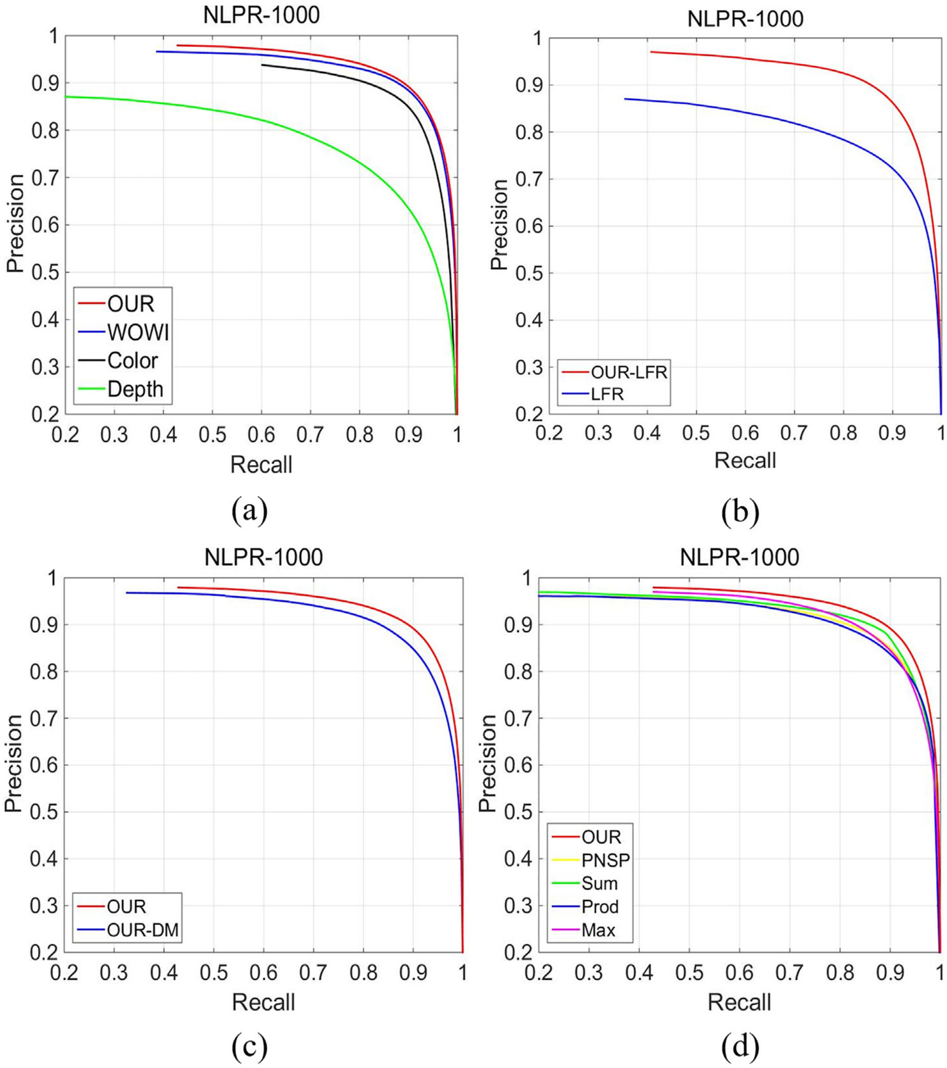


Fig. 8. Comparison of precision-recall (PR) curves among different model settings and fusion methods on the NLPR-1000 dataset.

is highlighted in our final saliency map generated by using our whole model. Therefore, the better visual quality of our final saliency map in Fig. 7(e) also demonstrates the rationality of our saliency model.

2) To analyze the effect of color saliency network, we replaced Amulet with LFR as the color saliency network and retained other parts as illustrated in Fig. 1. Let "LFR" denote the saliency maps generated by LFR, and "OUR-LFR" denote the saliency maps generated by our retrained model with LFR as the color saliency network. We can see from Table 4 and Fig. 8(b) that OUR-LFR consistently outperforms LFR in terms of all the three evaluation metrics. This demonstrates the robustness of our model to the use of existing different image saliency models, *e.g.* Amulet or LFR, as the color saliency network in our model, due that our model can exploit the proposed depth saliency network and saliency fusion network to make full use of depth information and elevate saliency detection performance.
3) To analyze the contribution of converting the depth map to the three-channel depth image, we used the depth map as the input to depth saliency network, and retrained our model to generate the final saliency maps, which are denoted as "OUR-DM". As show in Table 4 and Fig. 8(c), our results (OUR) with the three-channel depth image as the input to depth saliency network are consistently better than the results of OUR-DM in terms of all the three evaluation metrics, due that the three-channel depth image can enhance the depth discontinuities between salient object and background regions.
4) To analyze the contribution of saliency fusion network in our model, we also compared with the other four fusion methods in [42,43]. The four methods perform the fusion operation of product (Prod), sum (Sum), the parameterized normalization, sum and product (PNSP) and maximum (Max), respectively. We fused the depth saliency maps and the color saliency maps using the four fusion methods, respectively, and compared the fusion results with our results (OUR). We can see from Table 4 and Fig. 8(d) that the proposed saliency fusion network consistently outperforms the other four fusion methods.

4. Conclusion

In this paper, we propose an effective depth-aware saliency model using the three convolutional neural networks *i.e.* color saliency network, depth saliency network and saliency fusion network. The proposed deep saliency network adopts the weight sharing from the color saliency network, and the multi-layer feature pyramid structure, to improve the ability of depth feature extraction. Besides, the proposed saliency fusion network can effectively perform deep fusion between the color saliency map and depth saliency map to generate the high-quality final saliency map. Experimental results on five public datasets demonstrate the better saliency detection performance of our saliency model.

Conflict of interest

The authors declared that there is no conflict of interest.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 61771301 and 61801219.

References

[1] L. Itti, C. Koch, E. Niebur, A model of saliency-based visual attention for rapid scene analysis, IEEE Trans. Pattern Anal. Mach. Intell. 20 (1998) 1254–1259.
[2] Z. Liu, X. Zhang, S. Luo, O. Le Meur, Superpixel-based spatiotemporal saliency detection, IEEE Trans. Circuits Syst. Video Technol. 24 (2014) 1522–1540.
[3] Y. Xie, H. Lu, M. Yang, Bayesian saliency via low and mid level cues, IEEE Trans. Image Process. 24 (2013) 1689–1698.
[4] W. Zou, Z. Liu, K. Kpalma, J. Ronsin, Y. Zhao, N. Komodakis, Unsupervised joint salient region detection and object segmentation, IEEE Trans. Image Process. 24 (2015) 3858–3873.
[5] J. Yan, M. Zhu, H. Liu, Y. Liu, Visual saliency detection via sparsity pursuit, IEEE Signal Process. Lett. 17 (2010) 739–742.
[6] Z. Liu, W. Zou, O. Le Meur, Saliency tree: a novel saliency detection framework, IEEE Trans. Image Process. 23 (2014) 1937–1952.
[7] Y. Qin, H. Lu, Y. Xu, H. Wang, Saliency detection via cellular Automata, in: Proc. IEEE CVPR, Boston, MA, USA, 2015, pp. 110–119.
[8] J. Han, S. He, X. Qian, D. Wang, Lei. Guo, T. Li, An object-oriented visual saliency detection framework based on sparse coding representations, IEEE Trans. Circuits Syst. Video Technol. 23 (2013) 2009–2021.
[9] J. Han, R. Quan, D. Zhang, F. Nie, Robust object co-segmentation using background prior, IEEE Trans. Image Process. 27 (2018) 1639–1651.
[10] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, in: Proc. ICLR, San Diego, CA, USA, 2015, pp. 1–14.
[11] J. Zhang, P. Liu, F. Zhang, Q. Song, CloudNet: Ground-based cloud classification with deep convolutional neural network, Geophys. Res. Lett. 45 (2018) 8665–8672.
[12] R. Zhao, W. Ouyang, H. Li, X. Wang, Saliency detection by multi-context deep learning, in: Proc. IEEE CVPR, Boston, MA, USA, 2015, pp. 1265–1274.
[13] G. Li, Y. Yu, Visual saliency based on multiscale deep features, in: Proc. IEEE CVPR, Boston, MA, USA, 2015, pp. 5455–5463.
[14] L. Wang, H. Lu, X. Ruan, M.-H. Yang, Deep networks for saliency detection via local estimation and global search, in: Proc. IEEE CVPR, Boston, MA, USA, 2015, pp. 3183–3192.
[15] J. Guo, T. Ren, J. Bei, Salient object detection for RGB-D image via saliency evolution, in: Proc. IEEE ICME, Seattle, WA, USA, 2016, pp. 1–6.
[16] X. Wang, H. Ma, X. Chen, S. You, Edge preserving and multi-scale contextual neural network for salient object detection, IEEE Trans. Image Process. 27 (2018) 121–134.
[17] P. Zhang, D. Wang, H. Lu, H. Wang, X. Ruan, Amulet: Aggregating multi-level convolutional features for salient object detection, in: Proc. IEEE ICCV, Venice, Italy, 2017, pp. 202–211.
[18] G. Li, Y. Yu, Deep contrast learning for salient object detection, in: Proc. IEEE CVPR, Las Vegas, NV, USA, 2016, pp. 478–487.
[19] J. Han, D. Zhang, G. Cheng, N. Liu, D. Xu, Advanced deep-learning techniques for salient and category-specific object detection: a survey, IEEE Signal Process. Mag. 35 (2018) 84–100.
[20] P. Zhang, W. Liu, H. Lu, C. Shen, Salient object detection by lossless feature reflection, in: Proc. IJCAI, Morgan Kaufmann, Stockholm, Sweden, 2018, pp. 1–8.
[21] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, A. Borji, Detect globally, refine locally: a novel approach to saliency detection, in: Proc. IEEE CVPR, Salt Lake City, Utah, USA, 2018, pp. 3127–3135.
[22] R. Ju, Y. Liu, T. Ren, L. Ge, G. Wu, Depth-aware salient object detection using anisotropic center-surround difference, IEEE Signal Process. Lett. 38 (2015) 115–126.
[23] Y. Niu, Y. Geng, X. Li, F. Liu, Leveraging stereopsis for saliency analysis, in: Proc. IEEE CVPR, Providence, RI, USA, 2012, pp. 454–461.
[24] X. Fan, Z. Liu, G. Sun, Salient region detection for stereoscopic images, in: Proc. IEEE DSP, Hong Kong, China, 2014, pp. 454–458.
[25] H. Song, Z. Liu, H. Du, G. Sun, C. Bai, Saliency detection for RGBD images, in: Proc. ICIMCS, ACM, Zhangjiajie, China, 2015, pp. 240–243.
[26] Y. Fang, J. Wang, M. Narwaria, P. Le Callet, W. Lin, Saliency detection for stereoscopic images, IEEE Trans. Image Process. 23 (2014) 2625–2636.
[27] Y. Fang, C. Zhang, J. Li, J. Lei, M.P. Da Silva, P. Le Callet, Visual attention modeling for stereoscopic video: a benchmark and computational model, IEEE Trans. Image Process. 26 (2017) 4684–4696.
[28] H. Song, Z. Liu, H. Du, G. Sun, O. Le Meur, T. Ren, Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning, IEEE Trans. Image Process. 26 (2017) 4204–4216.
[29] G. Li, C. Zhu, A three-pathway psychobiological framework of salient object detection using stereoscopic technology, in: Proc. IEEE ICCVW, Venice, Italy, 2017, pp. 3008–3014.
[30] L. Qu, S. He, J. Zhang, J. Tian, Y. Tang, Q. Yang, RGBD salient object detection via deep fusion, IEEE Trans. Image Process. 26 (2017) 2274–2285.
[31] 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. Cybernetics 48 (2018) 3171–3183.
[32] C. Zhu, X. Cai, K. Huang, T.H. Li, G. Li, PDNet: Prior-model guided depth-enhanced network for salient object detection, in: Proc. IEEE ICME, San Diego, USA, 2018, pp. 1–6.
[33] C. Couprie, C. Farabet, L. Najman, Y. LeCun, Indoor semantic segmentation using depth information, in: Proc. ICLR, Scottsdale, AZ, USA, 2013, pp. 1–8.
[34] C. Hazirbas, L. Ma, C. Domokos, D. Cremers, FuseNet: Incorporating depth into semantic segmentation via fusion-based CNN architecture, in: Proc. ACCV, Springer, Taipei, Taiwan, 2016, pp. 213–228.

[35] H. Peng, B. Li, W. Xiong, W. Hu, R. Ji, RGBD salient object detection: A benchmark and algorithms, in: Proc. ECCV, Springer, Zurich, 2014, pp. 92–109.

[36] S. Gupta, R. Girshick, P. Arbeláez, J. Malik, Learning rich features from RGB-D images for object detection and segmentation, in: Proc. ECCV, Springer, Amsterdam, The Netherlands, 2014, pp. 345–360.

[37] S. Chopra, R. Hadsell, Y. LeCun, Learing a similarity metric discriminatively, with application to face verification, in: Proc. IEEE CVPR, San Diego, California, 2005, pp. 539–546.

[38] T. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, S. Belongie, Feature pyramid networks for object detection, in: Proc. IEEE CVPR, Honolulu, HI, USA, 2017, pp. 237–240.

[39] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, T. Darrell, Caffe: Convolutional architecture for fast feature embedding, in: Proc. ACM MM, Orlando, Florida, USA, 2014, pp. 675–678.

[40] Y. Cheng, H. Fu, X. Wei, J. Xiao, X. Cao, Depth enhanced saliency detection method, Proc. ICIMCS, Xiamen, China, 2014, article 23.

[41] N. Otsu, A threshold selection method from gray-level histograms, IEEE Trans. Syst. Man Cybern. 9 (1979) 62–66.

[42] A. Borji, M.M. Cheng, H. Jiang, J. Li, Salient object detection: a benchmark, IEEE Trans. Image Process. 24 (2015) 5706–5722.

[43] Y. Fang, W. Lin, Z. Chen, C.-M. Tsai, C.-W. Lin, A video saliency detection model in compressed domain, IEEE Trans. Circuits Syst. Video Technol. 24 (2014) 27–38.