Multi-level progressive parallel attention guided salient object detection for RGB-D images

Zhengyi Liu, Quntao Duan, Song Shi & Peng Zhao

The Visual Computer
International Journal of Computer Graphics

ISSN 0178-2789

Vis Comput
DOI 10.1007/s00371-020-01821-9

Springer

The Visual Computer
https://doi.org/10.1007/s00371-020-01821-9

ORIGINAL ARTICLE

Multi-level progressive parallel attention guided salient object detection for RGB-D images

Zhengyi Liu[1] · Quntao Duan[1] · Song Shi[1] · Peng Zhao[1]

Abstract
Detecting salient objects in RGB-D images attracts more and more attention in recent years. It benefits from the widespread use of depth sensors and can be applied in the comprehensive understanding of RGB-D images. Existing models focus on double-stream networks which transfer from color stream to depth stream, but depth stream with one channel information cannot learn the same feature as color stream with three channels information even if HHA representation is adopted. In our works, RGB-D four-channels input is chosen, and meanwhile, progressive parallel spatial and channel attention mechanisms are performed to improve feature representation. Spatial and channel attention can pay more attention on partial positions and channels in the image which show higher response to salient objects. Both attentive features are optimized by attentive feature from higher layer, respectively, and parallel fed into recurrent convolutional layer to generate side-output saliency maps guided by saliency map from higher layer. Last multi-level saliency maps are fused together from multi-scale perspective. Experiments on benchmark datasets demonstrate that parallel attention mechanism and progressive optimization operation play an important role in improving the accuracy of salient object detection, and our model outperforms state-of-the-art models in evaluation matrices.

Keywords Salient object detection · RGB-D image · Attention mechanism · Recurrent convolutional layer

1 Introduction

RGB-D sensors have been widely used in many human–machine interactive system. They can capture RGB-D images for better analyzing what they see. If RGB-D sensors can mimic human visual attention mechanism and recognize most attractive objects in their visual field, their intelligence will achieve impressive progress. The idea can come true by salient object detection for RGB-D images, which highlight salient objects in a scene by color and depth cues.

✉ Zhengyi Liu
liuzywen@ahu.edu.cn

Quntao Duan
984037793@qq.com

Song Shi
1271070920@qq.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

Early researches focus on extracting hand-crafted features, such as contrast prior [1–6], background prior [2,4,6,7], center prior and dark channel prior [8], to detect salient objects in RGB-D images. Nevertheless, it cannot always succeed in complex scenarios. Recently, convolutional neural networks (CNNs), which intelligently extract high-level and multi-scale complex representations from input images directly, have achieved superior performance in saliency detection [9–14]. However, different spatial positions and channels of the features from CNNs have different responses to salient objects or background in the image. Considering all spatial positions or all channels equally will lead to sub-optimal results. So attention mechanism [15,16], which assigns larger weights to spatial positions or channels which show higher response to salient objects, can extract effective attentive features and alleviate the distractions from background.

Saliency networks guided by attention mechanics for RGB images have been proposed in recent years. Literature [17–19] progressively refine global feature from top to down by attention mechanism. Attention mechanism is applied to progressively refine global saliency feature by shallower level

Published online: 18 February 2020

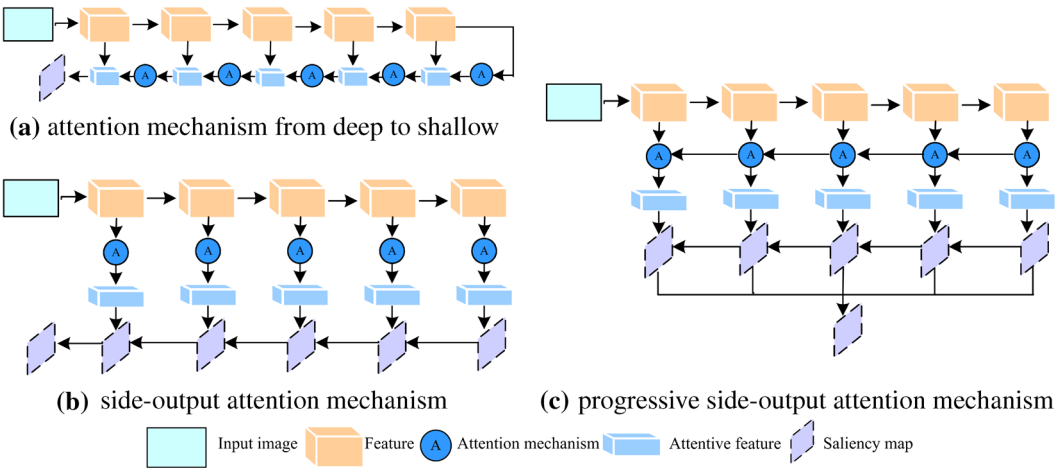


Fig. 1 Illustration of three attention mechanism architecture configurations: **a** attention mechanism architecture from deep to shallow. **b** Side-output attention mechanism architecture. **c** Progressive side-output attention mechanism architecture

feature, as shown in Fig. 1a. Literature [20–22] propose side-output attention mechanism to refine the feature from backbone network, and then, high-level semantic information is transmitted from deep to shallow for better locating salient objects in the shallowest layer. Attention mechanism is applied to side-output feature, and result saliency map is treated as the saliency map from the shallowest level attentive feature refined by deeper level attentive feature, as shown in Fig. 1b.

Our work is the first to design multi-level progressive parallel attention guided salient object detection model for RGB-D images. Attention mechanism is applied to side-output feature which is the same as side-output attention mechanism architecture (Fig. 1b), but its attentive feature is refined progressively from deep to shallow, as shown in Fig. 1c.

In contrast to RGB images, RGB-D image includes RGB color image and the paired depth image. How to use the depth image becomes an important problem in salient object detection. Recent researches almost adopt double-stream pattern [10,11,14,23–25], in which RGB and depth information are viewed with the equal importance and they are parallel processed. RGB image has three channels, and depth image has only one channel. Depth information with single channel is transformed to HHA encoding [26] (horizontal disparity, height above ground, angle with the direction of gravity) with three channels. But depth information provides auxiliary but not particularly trustworthy information [27], so double-stream pattern paying the equal importance to RGB and depth images is in more consideration of cross-modal transfer learning from RGB modal to depth modal. Meanwhile, references [28,29] point out that it is difficult to learn good depth-modal specific features by directly fine tuning RGB-specific model to depth-specific one. They argue that depth patterns in HHA encoding are typically smooth

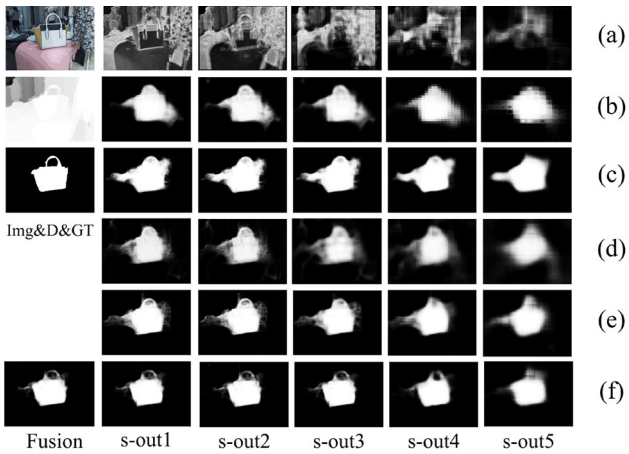


Fig. 2 Visual comparison of saliency map from each level before and after spatial attention and channel attention respective. The first column from top to down are input image, depth image, ground truth and result saliency map, and the rest of columns from left to right are five side-output layers, from top to down are: **a** saliency map from side-output, **b** spatial attention result, **c** progressive spatial attention result, **d** channel attention result, **e** progressive channel attention result and **f** optimized result by RCL

variations, contrasts and borders, but without textures and high-frequency patterns, and many low-level filters are either useless or ignored during fine tuning the network from RGB to HHA. What is more, double-stream mode needs two CNNs, which doubles the number of network parameters and computation cost. So, in our work, RGB-D four-channels input is adopted. Depth information is viewed as a channel input the same as Red, Green, Blue channels. More and more researches [30–32] use the single-stream network for fewer parameters and lower computational cost.

The multi-level features from backbone network are not the best, as shown in Fig. 2a, side-output saliency maps of the

network (the first row) are significant different from ground truth (the third row and first column). Then, attention mechanism plays an important role in extracting effective feature representation. Spatial attention mechanism can assign larger weights to positions which show higher response to salient objects in the image, and channel attention mechanism can focus more on the channels which show foreground regions. Such two attention mechanisms are parallelly performed on side-output feature from backbone network for better spatial and channel attentive feature, as shown in Fig. 2b, d, respectively. Meanwhile, progressive operation is performed for more accurate multi-level attentive feature. Attentive feature from deep layer can bridge more high-level semantic information to help the shallow layers get sharp object boundaries and also filter out noisy response in the background region, as shown in Fig. 2c, e, they are the progressive attentive results on the basis of (b, d). We can see that sharp object boundaries can be achieved by progressive optimization operation. Then, recurrent convolution layer (RCL) [33] is applied to fuse two progressive attentive features and the neighbor deeper saliency map to capture the local context information for generating attentive saliency map, as shown in Fig. 2f. Attentive saliency maps from different levels have different advantages, so the fusion of all the attentive maps is last adopted, as shown in the last row and first column of Fig. 2.

The contributions of this paper can be summarized as follows:

1. Spatial and channel attentive features are aggregated with spatial and channel attentive feature from higher layer, respectively, which is called progressive optimization operation, for better feature representation of side-output from backbone network.
2. Spatial and channel attention mechanisms are placed on the side-output of backbone network in a paralleled manner instead of the sequential manner. Spatial and channel attentive features can focus more on salient objects from two different perspectives. Further fusion with saliency map from higher layer by recurrent convolutional layer can extract enough comprehensive feature to generate the better side-output saliency maps.

2 Related works

2.1 Salient object detection for RGB-D images

Salient object detection for RGB-D images by CNN is attracting more and more interest, and becoming a main trend. Qu et al. [34] design convolutional neural network to fuse different low-level saliency cues into hierarchical features for automatically detecting salient objects in RGB-D images. The input of the network is feature vector instead of the image. It is different from end-to-end training network later.

Then, literature [10–12,14,23,25] pre-train RGB and depth modality independently and then stitch to optimize the eventual depth-induced saliency detection model. They belong to double-stream processing pattern. RGB information and depth information which is transformed to three-channel HHA representations [26] are parallel trained. But depth cue provides auxiliary but not particularly trustworthy information, depth stream supplementary to RGB stream is the better choice. Zhu et al. [9] propose an architecture which is composed of a master network for processing RGB cues, and a sub-network making full use of depth cues and incorporating depth-based features into the master network. Single-stream pattern which RGB and depth information are simultaneously processed provides the way to detect salient objects in the RGB-D images. More and more researches [30–32] use the single-stream network. At first, double stream needs two CNNs, which doubles the number of network parameters and computation cost; second, literature [28,29] point out that it is difficult to learn good depth-modal specific features by directly fine tuning RGB-specific model to depth-specific one. They argue that many low-level filters in HHA encoding are either useless or ignored during fine tuning the network from RGB to HHA. So double stream with transfer learning often takes bad effect. In the paper, single stream with RGB-D four-channels input is adopted.

2.2 Attention mechanism

Attention mechanisms have shown its efficiency in computer vision tasks, such as image captioning [35,36], visual question answering [37–39], pose estimation [40] and image classification [41].

Kuen et al. [42] first introduce attention mechanism to salient object detection. It generates initial saliency map by an end-to-end convolutional–deconvolutional network (CNN-DecNN) and then uses spatial transformer and recurrent network units to iteratively attend to selected image subregions to perform saliency refinement progressively. It considers only spatial attention achieved by arbitrary-size image subregions. Zhang et al. [17] introduce spatial and channel attention mechanisms to salient object detection for RGB images. Global feature can be progressively refined by channel attention and subsequent spatial attention from deep to shallow. Sun et al. [18] introduce self-attention mechanism to salient object detection for RGB images. Global feature can be progressively refined by self-attention, which assigns different position attention weights for salient objects and backgrounds, from deep to shallow. Liu et al. [19] introduce global spatial attention and local spatial attention mechanism to salient object detection for RGB images. Chen et

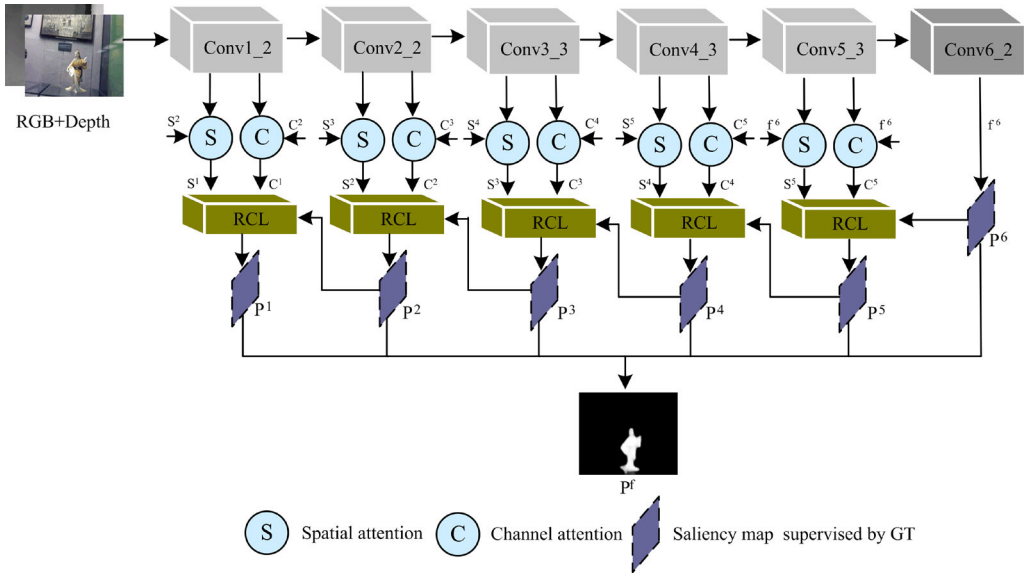


Fig. 3 Framework of multi-level progressive parallel attention guided salient object detection for RGB-D images

al. [20,21] propose reverse attention to guide side-output residual learning in a top-down manner. By erasing the current predicted salient regions from side-output features, the network can eventually explore the missing object parts and details which results in high resolution and accuracy. In fact, attention module in the paper is spatial attention weight expressed by saliency map from deeper layer. Zhang et al. [22] introduce a contextual attention module that can effectively guide the low-layer feature learning and force the backbone network to focus on the informative object regions with contextual pyramids.

The aforementioned works focus on salient object detection for RGB images by different attention mode. Some works belong to attention mechanism architecture from deep to shallow (Fig. 1a), and some works belong to side-output attention mechanism architecture (Fig. 1b). Our work pays more attention on RGB-D images with four-channels input, so multi-level progressive parallel attention mechanism is adopted to achieve more accurate attentive feature (Fig. 1c).

Meanwhile some works only consider spatial attention mechanism, and some works consider spatial and channel attention mechanism in the sequential manner. Our work delves into parallel spatial and channel attention mechanism and achieves the better results.

3 Proposed method

Due to the success of attention mechanism in computer vision tasks, we propose a multi-level progressive parallel attention guided salient object detection network for RGB-D images which generates parallel spatial and channel attentive features optimized by attentive feature from deeper layer, and then, a recurrent convolutional layer (RCL) [33] combined with two attentive feature and deeper saliency map is adopted for transferring global semantic information from top layer to shallower layers progressively, optimizing side-output feature further and generating side-output saliency maps. Finally, saliency maps from all the layers are aggregated together. The network is trained end-to-end, as shown in Fig. 3.

3.1 RGB-D feature extraction

VGG-16 network [43] is adopted as backbone network to extract features. RGB-D image demonstrates 3D real world, and depth information provides auxiliary but not particularly trustworthy information for salient object detection. So depth information in RGB-D image is regarded as attached cue in this paper. RGB image and paired depth image are wrapped to 224×224 size and fed into backbone network at the same time. In order to satisfy the need of four-channels input, the convolution kernel in the Conv1_1 of VGG-16 network is first modified from 64×3 × 3×3 to 64×3 × 3×4. Due to the nature of dense saliency detection problem, the last fully connected layers in VGG-16 network are replaced by two convolution layers named conv6_2. Its output feature f^6 performs deconvolutional operation to recover the size of image and then performs convolutional operation with 1×1 kernel to generate saliency map P^6.

3.2 Progressive attention mechanism

RGB-D four-channels input generates sub-optimal feature due to unreliable depth information; however, attention mechanism can focus on partial position or channels of the

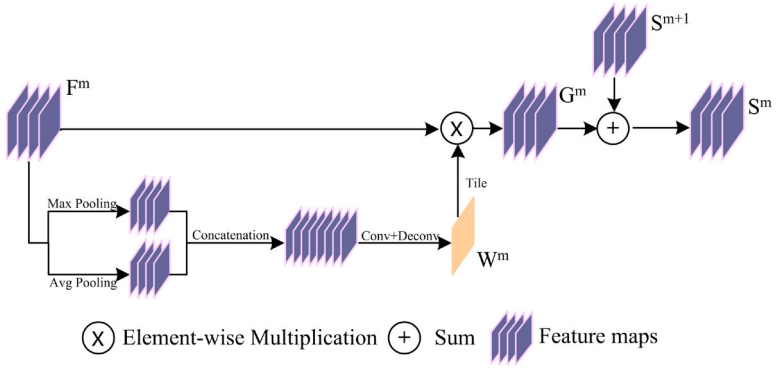


Fig. 4 Structure of progressive spatial attention mechanism

feature, which show higher response to salient object. So we attempt to apply attention mechanism to get more accurate feature. Inspired by [17], spatial and channel attention mechanisms are adopted, which assign larger weights to some positions or channels. But spatial and channel attention mechanisms are placed in a parallel manner instead of sequential manner which is described in [17]. Thus, comprehensive attentive features can be extracted from two different perspective. Meanwhile, multi-level progressive optimization operation is adopted to refine attentive feature further.

3.2.1 Progressive spatial attention mechanism

Spatial attention can focus more on the foreground region and reduce the interference of background noise rather than the equal consideration of all spatial locations, which helps to generate effective features for salient object detection.

Progressive spatial attention mechanism is proposed to utilize the inter-spatial relationship of features and spatial attentive feature from higher layer simultaneously, as shown in Fig. 4.

For hierarchical features $F^m(m = 1, \ldots, 5) \in \mathbb{R}^{H\times W\times C}$ from the side-output of backbone network, the average-pooling and max-pooling operations are utilized to retain more background information and more texture features [15]; then, their results are concatenated together to form the feature with 128 channels, and next the feature is squeezed across channel dimension into a spatial weight $W^m(m = 1, \ldots, 5) \in \mathbb{R}^{H\times W\times 1}$ by subsequent convolution and deconvolution operations, where spatial weight is denoted as:

$$W^m = Decov(Conv(Avg(F^m) \odot Max(F^m))) \tag{1}$$

where $Avg(\cdot)$ and $Max(\cdot)$ represent average-pooling and max-pooling operations, respectively, $\odot$ denotes concatenation operation, $Conv(\cdot)$ denotes two convolution operations, and $Deconv(\cdot)$ is up-sampling operation to keep the consistency with the size of feature F^m.

Next, the weights W^m are assigned to feature F^m and generate spatial attentive features $G^m \in \mathbb{R}^{H\times W\times C}$.

$$G^m = tile(W^m) \otimes F^m \tag{2}$$

where tile$(\cdot)$ operation denotes replicating W^m the same times as the number of the channels in the feature F^m, and $\otimes$ represents element-wise multiplication operation.

Last spatial attentive features S^{m+1} from the m+1th layer are aggregated into the mth layer spatial attentive features G^m and generate progressive spatial attentive features $S^m \in \mathbb{R}^{H\times W\times C}$ by normalization operation.

$$S^m = \frac{1}{1 + e^{-(G^m \oplus S^{m+1})}} \tag{3}$$

when $\oplus$ represents element-wise addition operation. When $m = 5$, $S^{m+1} = f^6$, which is the output feature from Conv6_2.

Thus, different weights are assigned to different spatial location of features by inter-spatial relation, and higher layer spatial attentive features are added for trustworthy progressive spatial attentive feature.

3.2.2 Progressive channel attention mechanism

Spatial attention mechanism focuses on "where" and assigns weights to features by considering space information, while channel attention mechanism pays more attention on response from different channels and express "what" is the most important. As a result, progressive channel attention mechanism is proposed to utilize the inter-channel dependencies of features and channel attentive feature from higher layer simultaneously, as shown in Fig. 5.

Hierarchical features $F^m(m = 1, \ldots, 5) \in \mathbb{R}^{H\times W\times C}$ from the side-output of backbone network are first passed through global average-pooling module, which aggregates the feature maps across spatial dimensions $H \times W$ to produce a channel vector $Z^m \in \mathbb{R}^{1\times 1\times C}$, which is defined as:

$$Z^m = \frac{\sum_{i=1}^{H}\sum_{j=1}^{W} F^m(i, j)}{H \times W} \tag{4}$$

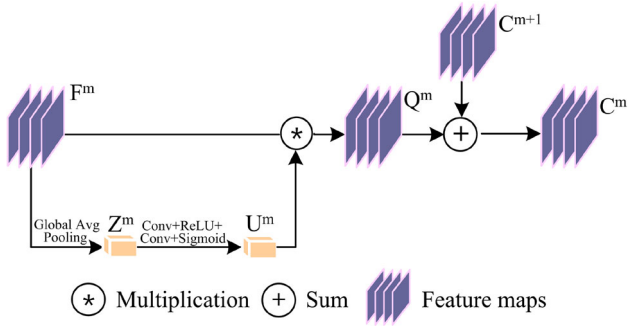


Fig. 5 Structure of progressive channel attention mechanism

Channel vector Z^m embeds the global distribution of channel-wise feature responses, enabling information from the global receptive field of the network to be leveraged by primitive side-out feature. It is then fed into two convolution layers with $1 * 1$ kernel to capture channel-wise dependencies. Next, a sigmoid activation layer used as a simple gating mechanism is adopted to gain the optimized channel weight $U^m \in \mathbb{R}^{1\times1\times C}$:

$$U^m = \frac{1}{1 + e^{-Conv(\delta(Conv(Z^m)))}} \tag{5}$$

where $Conv(\cdot)$ denotes convolution operation, and δ refers to the ReLU [44] function.

Next the weight U^m is assigned to feature F^m and generates channel attentive features $Q^m \in \mathbb{R}^{H\times W\times C}$.

$$Q^m = U^m \circledast F^m \tag{6}$$

where $\circledast$ represents multiplication operation.

Last channel attentive feature C^{m+1} from the m+1th layer is aggregated into the mth layer channel attentive features Q^m and generates progressive channel attentive features $C^m \in \mathbb{R}^{H\times W\times C}$ by a sigmoid activation layer.

$$C^m = \frac{1}{1 + e^{-(Q^m + C^{m+1})}} \tag{7}$$

when $m = 5$, $C^{m+1} = f^6$, which is the output feature from Conv6_2.

Thus, different weights are assigned to different channel of features by inter-channel relation, and channel attentive feature from higher layer is added for trustworthy progressive channel attentive feature.

3.3 Optimization by recurrent convolutional layer

Spatial and channel feature maps can strengthen spatial and channel information, but they miss subtle constructions and sharp boundaries. Recurrent convolutional layer (RCL) [33] is adopted to transfer global semantic information from top layer to shallower layers by combining the attentive cues and saliency map from higher layer.

For each RCL^m($m = 1, \ldots, 5$), as shown in Fig. 6, its input is composed of the output saliency maps P^{m+1}, the progressive spatial attentive feature S^m and the progressive channel attentive feature C^m. Then, three inputs are jointed together and are optimized by RCL to generate an optimized saliency map P^m in which the contextual information can be enhanced.

Each RCL^m with $T = 3$ can be unfolded to a feed-forward minute net with $T+1 = 4$ depth. The minute net has multiple paths from the input layer to the output layer by the multiple internal recurrent connection, which improves the ability of learning. Furthermore, the RCL units can accept larger and larger contexts by the expanding of the receptive field when the time step increases. Therefore, the RCL can help to incorporate local contexts, the spatial cues, the channel relationships and the high-level semantic information effectively.

The input $v_n^m(t)$ at time step t of the nth feature map in the RCL^m is given by:

$$v_n^m(t) = (w_n^{f^m})^T q^m + (w_n^{r^m})^T x^m(t-1) + b_n^m \tag{8}$$

where q^m and $x^m(t-1)$ are the feed-forward input from the previous layer and the recurrent input from the current layer at time step $t-1$, $w_n^{f^m}$ and $w_n^{r^m}$ denote feed-forward weights

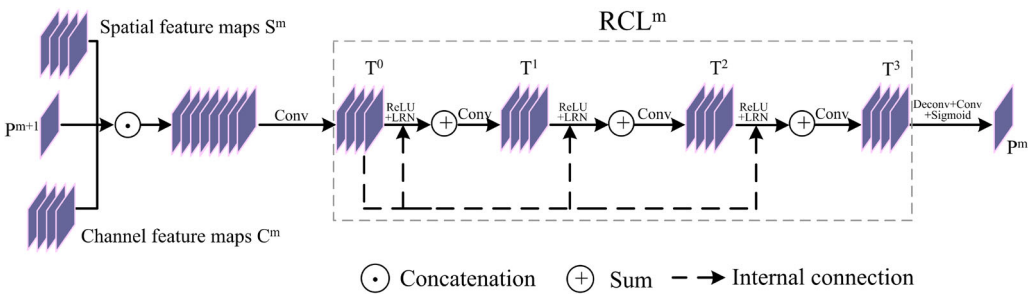


Fig. 6 Structure of recurrent convolutional layer optimization

and recurrent weights, respectively, b_n^m is the bias, and q^m is calculated by:

$$q^m = Conv(S^m \odot C^m \odot P^{m+1}) \tag{9}$$

where $\odot$ represents concatenation operation, and $Conv(\cdot)$ is the convolution operation with kernel size 3×3. For a unit on the nth feature map in an RCL^m, its state $x^m(t)$ at time step t is given by:

$$x^m(t) = g(f(v_n^m(t))) \tag{10}$$

where $f(\cdot)$ is the ReLU [33] function and $g(\cdot)$ is the local response normalization (LRN) function [33] to prevent the states from exploding.

To further restore the same image size as the input image, a deconvolutional layer is adopted. The subsequent convolutional layer and sigmoid activation function are applied to convert the feature maps to the saliency map $P^m(m = 1, \ldots, 5)$.

3.4 Fusion of hierarchical saliency maps

In order to combine multi-scale results, all the outputs $P^m(m = 1, \ldots, 6)$ are concatenated in the channel direction, and then, it is converted to the final saliency map P^f by a convolution operation, which is a weight-fused process combing the multi-scale local and global information of salient objects.

4 Experiments

4.1 Dataset

We evaluate the effectiveness of our model on three widely used public benchmark datasets.

NLPR1000 [45]. The NLPR RGB-D salient object detection dataset contains 1000 images captured by the Microsoft Kinect in various indoor and outdoor scenarios. It includes 11 types of indoor and outdoor scenes and more than 400 kinds of common objects under different illumination conditions.

NJU2000 [3]. The NJUDS2000 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.

STEREO [46]. The stereo dataset has 797 stereoscopic images. These images are mainly collected from the Internet and 3D movies. Depth images are generated by leveraging an optical method.

4.2 Evaluation matrices

PR curve The PR curves are plotted by comparing with the ground truth by setting a group of thresholds on the saliency maps to achieve binary masks.

MAE The MAE refers to the average pixel-wise error between the saliency map and ground truth.

F-measure The F-measure is computed by:

$$F_\beta = \frac{(1+\beta^2)\cdot \text{Precision}\cdot \text{Recall}}{\beta^2\cdot \text{Precision} + \text{Recall}} \tag{11}$$

where $\beta^2 = 0.3$ followed by [47].

E-measure E-measure proposed by [48] and [49] combines local information with global information. It is defined as follows:

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

where ϕ_{FM} denotes the enhanced-alignment matrix described in Fan et al. [49].

S-measure S-measure proposed by [48] and [49] solves the problem of structural measurement from the perspective of region-aware (S_r) and object-aware (S_o). S measure is defined as:

$$S_\lambda = \lambda * S_\text{o} + (1-\lambda) * S_\text{r} \tag{13}$$

where S_r and S_o are the object-aware and region-aware structural similarity, respectively, and λ is the balance parameter and is set as 0.5 in our experiment. In order to compare our method with other state-of-art methods fairly, all the evaluation use the same code[1] provided by [50].

4.3 Implementation details

The proposed model is implemented in python with Caffe toolbox. It is evaluated on a machine equipped with one GTX Titan-x GPUs (with 12G memory). The momentum, learning rate, weight decay and mini-batch size are set as 0.99, 1e−10, 0.0005 and 4, respectively. Data augmentation is adopted to solve data insufficiency problem. Each training image is horizontally and vertically flipped and cropped from top, down, left and right 1/10 image parts. So the training dataset is increased by 16 times, and meanwhile, the global semantic information is kept almost unchanged. The training process costs 7 h for 10 epochs. During testing, the model runs about 7 fps with 224×224 resolution.

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

Table 1 Ablation study

No.	Model	NLPR1000				NJU2000			
		$F\uparrow$	MAE$\downarrow$	$S\uparrow$	$E\uparrow$	$F\uparrow$	MAE$\downarrow$	$S\uparrow$	$E\uparrow$
1	*SNCN*	0.7760	0.0521	0.8715	0.8977	0.8134	0.0719	0.8600	0.8760
2	*SYCN*	0.7780	0.0591	0.8496	0.9023	0.7959	0.0801	0.8446	0.8798
3	*SYPCN*	0.8106	0.0420	0.8911	0.9188	0.8481	0.0569	0.8847	0.8944
4	*SNCY*	0.7678	0.0604	0.8749	0.8905	0.8030	0.0806	0.8685	0.8798
5	*SNCYP*	0.8168	0.0474	0.8674	0.9160	0.8437	0.0642	0.8667	0.8948
6	*SYPCYP*	**0.8279**	**0.0403**	**0.8967**	**0.9255**	**0.8639**	**0.0540**	**0.8900**	**0.9058**

The best result of each column is highlighted in bold. $\uparrow$ denotes the bigger value the better, and $\downarrow$ denotes the smaller value the better

Table 2 Comparison with sequential spatial and channel attention mechanism

No.	Model	NLPR1000				NJU2000			
		$F\uparrow$	MAE$\downarrow$	$S\uparrow$	$E\uparrow$	$F\uparrow$	MAE$\downarrow$	$S\uparrow$	$E\uparrow$
1	*spatial–channel*	0.7778	0.0540	0.8523	0.8970	0.8192	0.0804	0.8422	0.8783
2	*channel–spatial*	0.8141	0.0486	0.8617	0.9141	0.8398	0.0716	0.8549	0.8895
3	*parallel*	**0.8279**	**0.0403**	**0.8967**	**0.9255**	**0.8639**	**0.0540**	**0.8900**	**0.9058**

The best result of each column is highlighted in bold. $\uparrow$ denotes the bigger value the better, and $\downarrow$ denotes the smaller value the better

From results we can see that parallel model is superior than two sequential models, thus verifies our second contribution. Spatial and channel attentive features are optimized by spatial and channel attentive feature from higher layer, respectively. Both features and saliency map from higher layer are parallelly fed into RCL to get more comprehensive attentive feature, and further get the better side-output saliency maps.

4.4 Ablation study

In order to validate the effectiveness of our progressive attention mechanism, ablation studies are performed. Table 1 shows the ablation study results on NLPR1000 and NJU2000 datasets. Model $SNCN$ denotes that there is no spatial and channel attention mechanism. Model $SYCN$ denotes that there is only spatial attention mechanism. Model $SYPCN$ denotes that there is only progressive spatial attention mechanism. Model $SNCY$ denotes that there is only channel attention mechanism. Model $SNCYP$ denotes that there is only progressive channel attention mechanism. Model $SYPCYP$ denotes that there are both progressive channel and spatial attention mechanisms.

The 2th and 4th rows express spatial and channel attention mechanism influence on baseline model in the 1th row. From result, we can see that single spatial or channel attention mechanism does not show the impressive performance. The 3th and 5th rows express progressive optimization operation influence on the models in 2th and 4th row. From result, we can see that progressive optimization operation plays an impressive importance on saliency results and thus verifies our first contribution. Meanwhile, two progressive optimizations achieve the best performance, shown in the 6th row.

4.5 Comparison with sequential spatial and channel attention mechanism

In order to validate the effectiveness of our parallel attention mechanism, comparison with sequential spatial and channel attention mechanism [15,36] are performed. Sequential spatial and channel attention mechanism is classified into two categories. Spatial-channel model represents spatial first and channel second. Channel-spatial model represents channel first and spatial second.

Table 2 shows the comparison results on NLPR1000 and NJU2000 datasets.

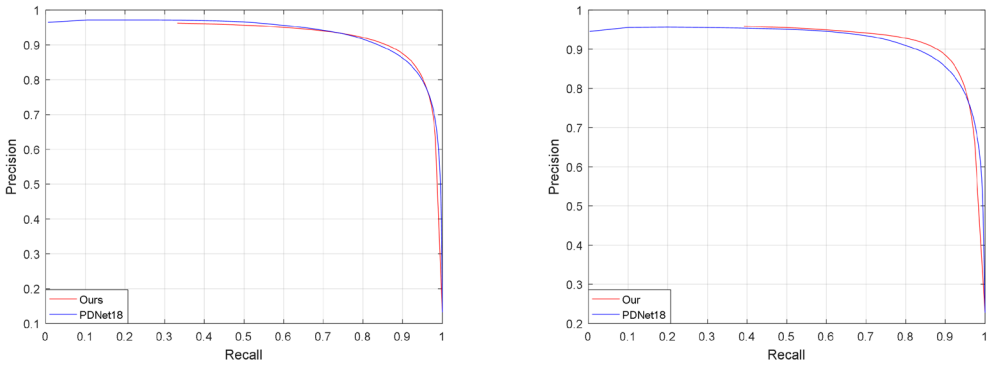
4.6 Comparison with other salient object detection models for RGB-D images

We compare our method with three traditional RGB-D saliency methods (ACSD [3], SE [51], LBE [6]) and seven RGB-D saliency methods by deep learning (DF [34], MLFN [10], CTMF [11], MMCI [13], PCFN [14], PDNet [9], TAN [52]). We use the codes provided by the authors to reproduce their experiments mainly for traditional methods or use the result saliency maps provided by the authors for deep learning based methods. Different training sets have an effect on result saliency maps. In order to fairly compare with other models, different training sets are adopted. One defined as $Ours\ (PDNet)$ is composed of 500 images in NLPR1000 and 1500 images in NJU2000, and the same with PDNet [9],

Table 3 F-measure, MAE, S-measure and E-measure comparisons of different models

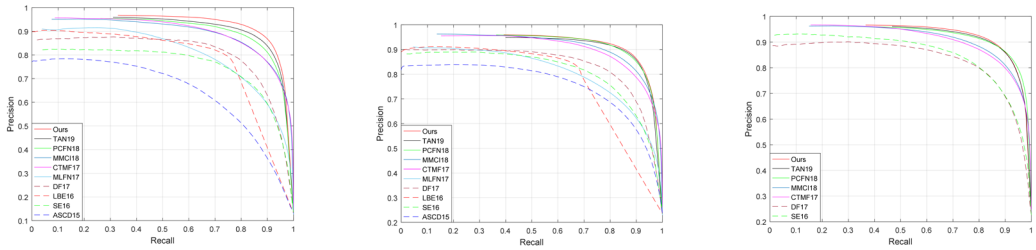
Data Set	NLPR1000				NJU2000				Stereo			
Metric	F-measure	MAE	S-measure	E-measure	F-measure	MAE	S-measure	E-measure	F-measure	MAE	S-measure	E-measure
ASCD15	0.5343	0.1787	0.6728	0.7417	0.6964	0.2021	0.6992	0.7863	–	–	–	–
SE16	0.6912	0.0913	0.7561	0.8385	0.7336	0.1687	0.6642	0.7722	0.7741	0.1452	0.7109	0.8308
LBE16	0.7355	0.0813	0.7619	0.8550	0.7400	0.1528	0.6952	0.7913	–	–	–	–
DF17	0.7348	0.0891	0.7909	0.8600	0.7703	0.1406	0.7596	0.8383	0.7650	0.1395	0.7664	0.8438
MLFN17	0.6413	0.0889	0.7900	0.8208	0.7046	0.1374	0.7709	0.8087	–	–	–	–
CTMF17	0.7234	0.0561	0.8599	0.8690	0.7875	0.0847	0.8490	0.8638	0.7859	0.0867	0.8529	0.8699
MMCI18	0.7299	0.0591	0.8557	0.8717	0.8122	0.0790	0.8581	0.8775	0.8120	0.0796	0.8559	0.8896
PCFN18	0.7948	0.0437	0.8736	0.9163	0.8440	0.0591	0.8770	0.8966	0.8450	0.0606	**0.8800**	0.9054
TAN19	0.7956	0.0410	0.8861	0.9161	0.8442	0.0605	0.8785	0.8932	0.8489	0.0591	0.8775	0.9108
Ours(PCFN)	**0.8322**	**0.0370**	**0.8939**	**0.9305**	**0.8540**	**0.0570**	**0.8816**	**0.9016**	**0.8513**	**0.0580**	0.8777	**0.9124**
PDNet18	0.7968	0.0501	0.8864	0.8987	0.8228	0.0709	0.8770	0.8882	–	–	–	–
Ours(PDNet)	**0.8279**	**0.0403**	**0.8967**	**0.9255**	**0.8639**	**0.0540**	**0.8900**	**0.9058**	**0.8529**	**0.0549**	**0.8873**	**0.9108**

The best result of each column is highlighted in bold



(a) NLPR1000 dataset **(b)** NJU2000 dataset

Fig. 7 P–R curves comparison of different models on NLPR1000 and NJU2000 dataset based on $Ours\ (PDNet)$ training set



(a) NLPR1000 dataset **(b)** NJU2000 dataset **(c)** Stereo dataset

Fig. 8 P–R curves comparison of different models on NLPR1000, NJU2000 and stereo dataset based on $Ours\ (PCFN)$ training set

and the other defined as $Ours\ (PCFN)$ is composed of 650 images in NLPR1000 and 1400 images in NJU2000, and the same with MLFN [10], CTMF [11], MMCI [13], PCFN [14] and TAN [52]. The training set of PDNet [9] is different from others, so it is compared alone, as shown in the last two rows in Table 3 and Fig. 7. The rest comparison results are shown in the rest rows in Table 3 and Fig. 8. Some data are missing because the author does not provide result saliency maps.

Although there is intersection among PDNet [9], PCFN [14], TAN [52] and our model in PR curve on some datasets, it wins them in more comprehensive evaluation matrices F-measure, MAE, S-measure and E-measure, as shown in Table 3. Figure 9 provides a visual comparison of our model with the above-mentioned models. It can be observed that our model produces fine detail and highlights the attention-grabbing salient region.

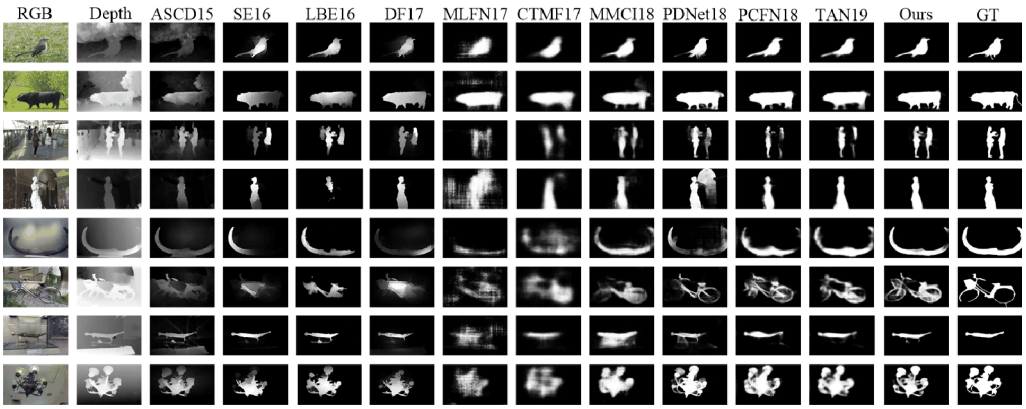


Fig. 9 Visual comparisons of different models

5 Conclusion

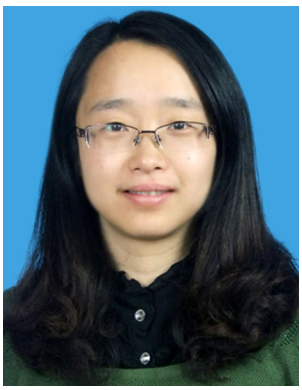
In this paper, we proposed a multi-level progressive parallel attention guided network for salient object detection in RGB-D images, in which spatial and channel attentive feature is optimized by attentive feature from deeper layer, and parallelly fed into recurrent convolutional layer for superior multi-level feature representation. Influenced by these two contributions, our model outperforms other state-of-the-art models in three benchmark datasets. Depth sensors equipped with our saliency detection model can improve its ability both for detecting depth value and analyzing visual scene and thus enhance the ability to sense the surrounding environment.

Acknowledgements 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 code of the evaluation metrics. We 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. Wu, P., Duan, L., Kong, L.: RGB-D salient object detection via feature fusion and multi-scale enhancement. In: CCF Chinese Conference on Computer Vision. Springer, pp. 359–368 (2015)
2. Ren, J., Gong, X., Yu, L., Zhou, W., Ying Yang, M.: Exploiting global priors for RGB-D saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 25–32 (2015)
3. Ju, R., Ge, L., Geng, W., Ren, T., Wu, G.: Depth saliency based on anisotropic center-surround difference. In: IEEE International Conference on Image Processing, pp. 1115–1119 (2015)
4. Jiang, L., Koch, A., Zell, A.: Salient regions detection for indoor robots using RGB-D data. In: 2015 IEEE International Conference on Robotics and Automation (ICRA), pp. 1323–1328. IEEE (2015)
5. Guo, J., Ren, T., Bei, J., Zhu, Y.: Salient object detection in RGB-D image based on saliency fusion and propagation. In: Proceedings of the 7th International Conference on Internet Multimedia Computing and Service, p. 59. ACM (2015)
6. Feng, D., Barnes, N., You, S., McCarthy, C.: Local background enclosure for RGB-D salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2343–2350 (2016)
7. Lin, H., Lin, C., Zhao, Y., Wang, A.: 3D saliency detection based on background detection. J. Vis. Commun. Image Represent. **48**, 238–253 (2017)
8. Zhu, C., Zhang, W., Li, T.H., Li, G.: Exploiting the value of the center-dark channel prior for salient object detection. arXiv preprint arXiv:1805.05132
9. Zhu, C., Cai, X., Huang, K., Li, T.H., Li, G.: PDNet: prior-model guided depth-enhanced network for salient object detection. arXiv preprint arXiv:1803.08636
10. Chen, H., Li, Y., Su, D.: RGB-D saliency detection by multi-stream late fusion network. In: International Conference on Computer Vision Systems, pp. 459–468. Springer (2017)
11. Han, J., Hao, C., Liu, N., Yan, C., Li, X.: CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. IEEE Trans. Cybern. **48**(99), 1–13 (2017)
12. Chen, H., Li, Y.-F., Su, D.: M3net: multi-scale multi-path multi-modal fusion network and example application to RGB-D salient object detection. In: Intelligent Robots and Systems (IROS), pp. 4911–4916. IEEE (2017)
13. Chen, H., Li, Y., Su, D.: Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. Pattern Recognit. **86**, 376–385 (2019)
14. Chen, H., Li, Y.: 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 (2018)
15. Woo, S., Park, J., Lee, J.-Y., So Kweon, I.: CBAM: convolutional block attention module. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 3–19 (2018)
16. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. arXiv preprint arXiv:1709.01507
17. Zhang, X., Wang, T., Qi, J., Lu, H., Wang, G.: Progressive attention guided recurrent network for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 714–722 (2018)
18. Sun, F., Li, W., Guan, Y.: Self-attention recurrent network for saliency detection. Multimed. Tools Appl. **78**, 1–15 (2018)
19. Liu, N., Han, J., Yang, M.-H.: Picanet: Learning pixel-wise contextual attention for saliency detection. arXiv preprint arXiv:1708.06433

20. Chen, S., Tan, X., Wang, B., Hu, X.: Reverse attention for salient object detection. arXiv preprint arXiv:1807.09940
21. Chen, S., Wang, B., Tan, X., Hu, X.: Embedding attention and residual network for accurate salient object detection. IEEE Trans. Cybern (2018). https://doi.org/10.1109/TCYB.2018.2879859
22. Zhang, P., Wang, L., Wang, D., Lu, H., Shen, C.: Agile amulet: real-time salient object detection with contextual attention. arXiv preprint arXiv:1802.06960
23. Chen, H., Li, Y., Su, D.: RGB-D salient object detection based on discriminative cross-modal transfer learning. arXiv preprint arXiv:1703.00122
24. Chen, H., Li, Y., Su, D.: Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. Pattern Recognit. **86**, 376–385 (2019)
25. Wang, S., Zhou, Z., Jin, W., Qu, H.: Visual saliency detection for RGB-D images under a bayesian framework. IPSJ Trans. Comput. Vis. Appl. **10**(1), 1 (2018)
26. Gupta, S., Girshick, R., Arbeláez, P., Malik, J.: Learning rich features from RGB-D images for object detection and segmentation. In: European Conference on Computer Vision, pp. 345–360. Springer (2014)
27. Cong, R., Lei, J., Zhang, C., Huang, Q., Cao, X., Hou, C.: Saliency detection for stereoscopic images based on depth confidence analysis and multiple cues fusion. IEEE Signal Process. Lett. **23**(6), 819–823 (2016)
28. Du, D., Xu, X., Ren, T., Wu, G.: Depth images could tell us more: enhancing depth discriminability for RGB-D scene recognition. In: 2018 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2018)
29. Song, X., Herranz, L., Jiang, S.: Depth CNNs for RGB-D scene recognition: learning from scratch better than transferring from RGB-CNNs. In: AAAI, pp. 4271–4277 (2017)
30. Liu, Z., Shi, S., Duan, Q., Zhang, W., Zhao, P.: Salient object detection for RGB-D image by single stream recurrent convolution neural network. Neurocomputing **363**, 46–57 (2019)
31. Huang, P., Shen, C.-H., Hsiao, H.-F.: RGBD salient object detection using spatially coherent deep learning framework. In: 2018 IEEE 23rd International Conference on Digital Signal Processing (DSP), pp. 1–5. IEEE (2018)
32. Fan, D.-P., Lin, Z., Zhao, J.-X., Liu, Y., Zhang, Z., Hou, Q., Zhu, M., Cheng, M.-M.: Rethinking RGB-D salient object detection: models, datasets, and large-scale benchmarks. arXiv preprint arXiv:1907.06781
33. Liang, M., Hu, X.: Recurrent convolutional neural network for object recognition. In: Computer Vision and Pattern Recognition, pp. 3367–3375 (2015)
34. Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., Yang, Q.: Rgbd salient object detection via deep fusion. IEEE Trans. Image Process. **26**(5), 2274–2285 (2017)
35. Xu, K., Ba, J., Kiros, R., Cho, K., Courville, A., Salakhudinov, R., Zemel, R., Bengio, Y.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
36. Chen, L., Zhang, H., Xiao, J., Nie, L., Shao, J., Liu, W., Chua, T.-S.: Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6298–6306. IEEE (2017)
37. Xu, H., Saenko, K.: Ask, attend and answer: exploring question-guided spatial attention for visual question answering. In: European Conference on Computer Vision, pp. 451–466. Springer (2016)
38. Yang, Z., He, X., Gao, J., Deng, L., Smola, A.: Stacked attention networks for image question answering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 21–29 (2016)
39. Zhu, Y., Groth, O., Bernstein, M., Fei-Fei, L.: Visual7w: grounded question answering in images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4995–5004 (2016)
40. Chu, X., Yang, W., Ouyang, W., Ma, C., Yuille, A.L., Wang, X.: Multi-context attention for human pose estimation. arXiv preprint arXiv:1702.07432 1(2)
41. Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X.: Residual attention network for image classification. arXiv preprint arXiv:1704.06904
42. Kuen, J., Wang, Z., Wang, G.: Recurrent attentional networks for saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3668–3677 (2016)
43. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Comput. Sci (2014). arXiv preprint arXiv:1409.1556
44. Nair, V., Hinton, G.E.: Rectified linear units improve restricted Boltzmann machines. In: Proceedings of the 27th International Conference on Machine Learning (ICML-10), pp. 807–814 (2010)
45. Peng, H., Li, B., Xiong, W., Hu, W., Ji, R.: RGBD salient object detection: a benchmark and algorithms. In: European Conference on Computer Vision, pp. 92–109. Springer (2014)
46. Niu, Y., Geng, Y., Li, X., Liu, F.: Leveraging stereopsis for saliency analysis. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 454–461. IEEE (2012)
47. Martin, D.R., Fowlkes, C.C., Malik, J.: Learning to detect natural image boundaries using local brightness, color, and texture cues. IEEE Trans. Pattern Anal. Mach. Intell. **26**(5), 530–549 (2004)
48. Fan, D.-P., Cheng, M.-M., Liu, Y., Li, T., Borji, A.: Structure-measure: a new way to evaluate foreground maps. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4558–4567 (2017)
49. Fan, D.-P., Gong, C., Cao, Y., Ren, B., Cheng, M.-M., Borji, A.: Enhanced-alignment measure for binary foreground map evaluation. arXiv preprint arXiv:1805.10421
50. Fan, D.-P., Cheng, M.-M., Liu, J.-J., Gao, S.-H., Hou, Q., Borji, A.: Salient objects in clutter: bringing salient object detection to the foreground. In: European Conference on Computer Vision, pp. 196–212. Springer (2018)
51. Guo, J., Ren, T., Bei, J.: Salient object detection for RGB-D image via saliency evolution. In: Multimedia and Expo (ICME), pp. 1–6. IEEE (2016)
52. Chen, H., Li, Y.: Three-stream attention-aware network for RGB-D salient object detection. IEEE Trans. Image Process. **28**(6), 2825–2835 (2019)



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

Quntao Duan is a M.S. candidate of Anhui University. She received her B.S. from Jiangxi agricultural University, China, in 2017. Her research interests include image and video processing and computer vision.

Song Shi is a M.S. candidate of Anhui University. He received his B.S. from Anhui University, China, in 2016. His research interests include image and video processing and computer vision.

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