

# Self-Attention Recurrent Network for Saliency Detection

Fengdong Sun · Wenhui Li · Yuanyuan Guan

Received: date / Accepted: date

**Abstract** Feature maps in deep neural network generally contain different semantics. Existing methods often omit their characteristics that may lead to sub-optimal results. In this paper, we propose a novel end-to-end deep saliency network which could effectively utilize multi-scale feature maps according to their characteristics. Shallow layers often contain more local information, and deep layers have advantages in global semantics. Therefore, the network generates elaborate saliency maps by enhancing local and global information of feature maps in different layers. On one hand, local information of shallow layers is enhanced by a recurrent structure which shared convolution kernel at different time steps. On the other hand, global information of deep layers is utilized by a self-attention module, which generates different attention weights for salient objects and backgrounds thus achieve better performance. Experimental results on four widely used datasets demonstrate that our method has advantages in performance over existing algorithms.

**Keywords** Saliency Detection · Recurrent Convolutional Layer · Self Attention Module

---

This work was supported by the Science and Technology Development Plan of Jilin Province under Grant 20170204020GX, the National Science Foundation of China under Grant U1564211.

---

Fengdong Sun  
College of Computer Science and Technology, Jilin University, Changchun, China, 130012

Wenhui Li  
College of Computer Science and Technology, Jilin University, Changchun, China, 130012  
E-mail: liwh@jlu.edu.cn

Yuanyuan Guan  
College of Computer Science and Technology, Jilin University, Changchun, China, 130012

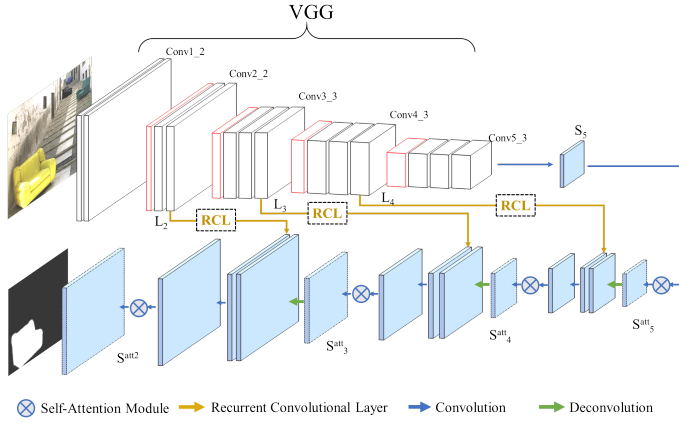
## 1 Introduction

Saliency detection, which locates the regions most attracting human beings, is an important branch of image processing. The goal of saliency detection is to find the most distinctive regions in an given image. Saliency detection has attracted widespread attentions owing to its widely application and high research values. Therefore, many efficient and robust saliency detection methods are developed recently. Saliency detection methods can be used as image preprocessing, due to the valuable semantic information that are contained by salient regions. The performance of many fields in computer vision and image processing can be enhanced by employing saliency detection, such as content-aware image editing[5,40], image compression[8], visual tracking[4], person re-identification [3,35,37], image retrieval[6,28], and video summarization[22,36]. However, improving the accuracy of saliency detection, especially in a clutter, is still a huge challenge.

The early saliency detection methods are generally inspired by the visual attention model proposed by Itti[13]. This kind of method usually extracts features manually, and calculates the visual contrast of each region via these handcrafted features. These methods follow a principle that the most salient regions have the highest visual contrast. Therefore global contrast and local contrast, which are two common measurements, are developed to simulate the visual contrast, and many saliency features are exploited based on the global and local contrasts in previous studies. However, the accuracy of methods based on handcrafted features is not satisfactory in a clutter background.

To obtain reliable and robust results, machine learning algorithms are developed to enhance the performance of saliency detection methods [33]. In the beginning, machine learning algorithms are employed to detect salient objects based on different handcrafted features. However, the deficiencies of handcrafted features could not be eliminated by this way. With the purpose of overcoming the drawback of handcrafted features, more methods based on deep convolutional neural networks (CNNs) are emerging. Depending on the learning ability of CNNs, the accuracy of saliency detection has been improved significantly. And end-to-end convolutional neural network could directly generate the salient maps without any manual operations so that it can make up the deficiencies of handcrafted features. The end-to-end network is generally composed of convolution operations, pooling operations, etc. The saliency features are generated in the process of convolution operations. Due to different sizes of receptive fields, shallow layers often contain more local information, and deep layers contain more global information. Therefore, how to utilize the convolutional information of different layers is still a key problem. Shallow layers contain plentiful local saliency information, there are lack of effective methods to enhance and take advantage of the local information. Moreover, deep layers contain plentiful global saliency information, which is need to be enhanced to highlight salient regions and suppress interference of background.

To overcome the aforementioned issues, we propose a novel end-to-end convolutional neural networks structure, which combined self-attention mechanism and recurrent convolutional layers (RCL) to enhance global and local saliency information. Deep network structure we proposed in this paper is composed of two subnet-



**Fig. 1** The entire network structure proposed in this paper.

works as shown in Figure 1. One subnetwork is used to extract feature maps based on VGG16[23]. The other subnetwork, called attentional recurrent network (ARN), is used to fuse different feature maps generated by VGG16. In the ARN subnetwork, RCL is used to receive the feature maps from shallow layers, and enhance the local saliency information in these feature maps with a shared weight recurrent structures. Moreover, an attention mechanism called self-attention is used to handle the feature maps in deep layers. Self-attention mechanism is used to obtain attentional weights, which are assigned more to salient regions for improving global saliency perception ability of ARN. The network proposed in this paper can capture subtle visual contrast for saliency detection, and generate delicate saliency maps. Experimental results demonstrate that our method have a better performance than 7 exact algorithms on four open datasets. In summary, contributions of this paper are as followings:

1. We propose a novel end-to-end deep convolutional neural network for robust saliency detection. The network consists of two parts. One part based on VGG16 is used to collect multi-scale features which contain visual contrast information, the other part called ARN is used to generate subtle and robust saliency maps.
2. A recurrent structure called RCL is utilized to handle feature maps in shallow layers. RCL enhances the local saliency information of these feature map by a shared recurrent convolutional operations with different time steps.
3. Self-attention is utilized to enhance the global saliency information. This kind of attention generates attentional weights based on input feature maps, and gives salient regions more weights to obtain more exact results.

## 2 Related Work

In the last decades, there emerged many saliency detection methods either in bottom-up or top-down. Bottom-up methods generally used the low-level features such as

color, texture, and edge information to generate final saliency maps. Top-down methods depended on high-level knowledge and generally detected salient objects by machine learning algorithm. Early saliency detection methods were common to use bottom-up methods with low-level features. The model proposed by Itti et al.[13], which was inspired by the theory Koch and Ullman [14] introduced, used center-surrounded features to measure visual difference in images. Cheng et al.[7] proposed a saliency detection method based on global color information. Histogram-based contrast was calculated by the distances between different colors. Region contrast was obtained by calculating the distances between regional color histograms globally. This method simplified colors in input images and generated saliency maps by histogram-based contrast and region contrast. Zhu [46] proposed a novel background prior, which evaluated the probability of regions belonging to background. And a saliency detection method based this background prior was put forward. Many other bottom-up methods were developed using different low-level features [42,46], and some methods employed supervised algorithms to improve performance, such as support vector machine (SVM) [34], hidden Markov model (HMM) [12] and conditional random fields (CRF) [39]. And clustering algorithms are very common in the traditional methods. [32,30,31,29]

Recently, more works based on deep learning were emerging. Among them, CNNs were preferred by many researchers because of its outstanding performance on image processing. Different kinds of CNNs was proposed to detect salient objects. A two-stream deep contrast network was presented by Li et al.[17], which consisted of two components, a pixel-level fully convolutional stream and a segment-wise spatial pooling stream. The two streams were used to generate subtle pixel-level saliency maps, and reduce the redundancy in computation and storage. Zhang et al.[44] proposed combining multi-level convolutional features and contextual attention module to generate saliency map. Some researches utilized the side semantic information of network to increase the accuracy of saliency detection. Thus, Hou et al.[11] proposed to utilize the side outputs information to enhance semantic information in deeper layer. Furthermore, recurrent structure in deep network has been proved that could help network to refine the semantics, and many different kinds of recurrent structure have been employed for saliency detection. Wang et al.[24] proposed a novel recurrent fully convolutional network to encode high level semantic features for saliency detection. The model refined outputs by the same network structure at different time steps for elaborate saliency maps. Liu et al.[20] proposed a saliency model called DHSnet which included two subnetworks. The first subnetwork based on VGG16 aimed to generate a coarse saliency map. The second subnetwork called hierarchical recurrent convolutional neural network was used to improve the coarse saliency map in details. A subtle saliency map was produced by the two streams jointly. As the researches about neural network moving along, many researchers paid more attention to employ visual attention into CNNs. The attention mechanism could assign different weights to feature maps according to their semantics, and helped network attach more importance to high weight foreground regions. Zhang et al.[45] proposed a multi-path recurrent module, which transferred global information from deep layers to shallower, to enhance the global semantics in shallower layers. And attentional features were used to alleviate distraction of background. Kuen et al.[16] utilized spatial

attention transforms to generate robust attention features, which were used to refine the final saliency maps.

### 3 Proposed Method

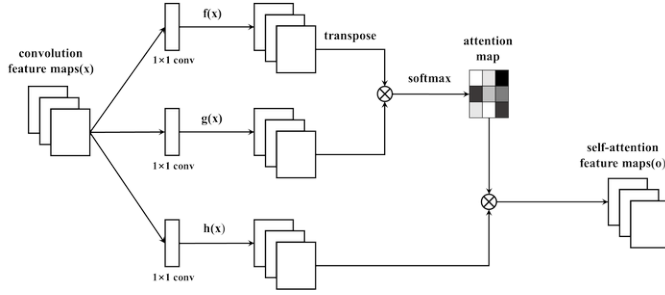
We introduce a deep contrast network for pixel-level saliency detection. A VGG network pre-trained on imageNet is used as a feature extraction network to extract contrast feature maps which contain abundant local and global semantics. And a novel deep recurrent network called ARN are employed to integrate these feature maps and generate pixel-level saliency map, wherein attention mechanisms are deployed to suppress the interference of backgrounds.

#### 3.1 Overall structure

In this section, we will introduce the overall structure of the proposed network structure. We adopt a fully convolutional framework which is efficient for convolution. As illustrated in Figure 1, the whole network could be divided into two parts. The first is the feature maps extraction part, which is used to generate multi-scale local feature maps and global feature maps. The other part is the aggregated part called attentional recurrent network, which is used to aggregate these feature maps generated by the extraction part and output the final saliency maps.

In the feature maps extraction part, we used a convolutional network called VGG16 which is popular in image classification and saliency detection. The network could generate reliable and robust feature maps for saliency detection. The side outputs and the last convolutional feature map of VGG16 are utilized to generate multi-scale features in this paper. In Section 3.2 we will elaborate the structure and useful outputs of the feature extraction network. As figure 1 illustrated, the outputs of feature extraction network could be denoted as  $L_2, L_3, L_4$  and  $S_5$ . Among them,  $L_2, L_3, L_4$  are the side outputs with different sizes, and  $S_5$  is the feature maps generated by forward propagation. These feature maps are fed into the attentional recurrent network to aggregate multi-scale semantics and generate pixel-wise saliency maps.

The feature maps in shallow layers, such as  $L_2, L_3, L_4$ , contain more local saliency information. And those in deep layers such as  $S_5$  contain more global saliency information. Therefore, we develop an attentional recurrent network to generate final saliency maps by these feature maps with different semantics. The ARN includes a series of transposed convolution layers to restore the size of feature maps, and several convolution operations to obtain more subtle global information. Self-attention module and recurrent convolutional layers in ARN are used to alleviate distractions and enhance local information respectively, which are introduced in detail in section 3.3 and section 3.4.  $L_2, L_3, L_4$  would be fed into the RCL unit to generate the enhanced features. The attentional feature maps  $S_5^{att}$  is generated by  $S_5$  after self-attention module.  $S_5^{att}$  is upsampled by transposed convolution operations and concatenated with the RCL enhanced side-outputs which have the same size. Then, the next stage attention feature maps  $S_4^{att}$  are generated after convolutional operations and self-attention



**Fig. 2** Self-attention module.

module. The final saliency map could be obtained by the last attention feature maps  $S_2^{att}$ .

### 3.2 Feature maps extraction

In this section, we introduce the network we used to extract feature maps. For an input image, we resize the image to size  $224 \times 224$  and feed it into VGG net which consists of 16 convolutional layers. These feature maps are generated by the hidden convolutional layers in VGG16 [23]. The last layer have a small size  $14 \times 14$ . The feature maps in this layer have been convoluted multiple times to the smallest size, and contain abundant global information. The feature maps in shallow layer have a larger size which means they convey more local information. (For presentation purposes, we denote Conv to represent the convolutional layers in VGG16, and Conv2\_1 represents the first sublayer in the second group of convolutional layers.) Therefore, we use the first 13 layers in VGG16 network to extract feature maps for local information, and use the last layer for global information. Because of deeper layers containing more semantic information, we utilize the last convolutional layer in every group, i.e., Conv2\_2, Conv3\_3, and Conv4\_3 as side outputs which are fed into RCL unit. Conv5\_3 is conducted by transposed convolutions to recover the size. Among these layers, Conv1\_2 has the same size  $224 \times 224$  with input images, and the size of other feature maps are halved from top layer group, i.e.  $112 \times 112$ ,  $56 \times 56$ ,  $28 \times 28$  and  $14 \times 14$ . From these layers, we could obtain multi-scale feature maps for saliency detection.

These feature maps contain different image semantic features including saliency cues. The layers, wherein the feature maps are generated, determined the size of receptive fields of each feature map. Receptive fields refer to the size of input image corresponding to a node on feature maps, and deeper layers generate small size feature maps with large receptive fields. The differences of receptive fields in size determine that each feature map contains different semantic information. Due to different scales and channels of these feature maps, an aggregated structure is necessary to integrate these feature maps and generate elaborate saliency maps.

### 3.3 Self-attention module

In general, salient objects only have a close relationship with partial regions of inputs. Attention mechanism is used to give the feature maps different weights to highlight the salient regions and alleviate the interference of backgrounds. We use a self-attention module in this paper, which calculates the weight of a position in a layer by attending to all feature maps within the same layer. To our best knowledge, the self-attention has not yet been explored in the saliency detection. In this section, we introduce the self-attention module to our ARN network, enabling ARN to efficiently improve the ability of exploring the important regions in layers. The self-attention module, which is introduced in Ref. [41], is shown as Figure 2.

The feature maps from the previous layers is  $x$ , which shape is  $\{W, H, C\}$ . These feature maps are first convoluted with  $1 \times 1$  kernel to generate attention features.

$$f(x) = W_f * x, g(x) = W_g * x \quad (1)$$

where  $*$  denotes convolution operation, and  $W_f$  and  $W_g$  are the convolution kernels with  $C_1$  channels. Therefore, the attention features which integrate different information of all channels could be represented as  $f(x), g(x) \in \mathbb{R}^{W \times H \times C_1}$ . Then the attention map can be calculated as Eq. 2

$$\beta = \frac{\exp(s)}{\sum_{i=1}^N \exp(s)} \quad (2)$$

where  $s = f(x)^T g(x)$ , in which  $f(x)$  and  $g(x)$  are reshaped to  $\{C_1, W \times H\}$  and  $\beta$  is the attention map which indicates the weights of all positions in feature maps. The shape of  $\beta$  is  $\{W \times H, W \times H\}$ , and  $C_1$  is set to  $C/8$  following the setting of Ref.[41]. Therefore, the weighted attention output could be represented as Eq. 3

$$o = \beta \otimes h(x) \quad (3)$$

where  $h(x) = W_h * x$ , in which the shape of  $W_h$  is  $\{W \times H, C\}$ .  $\otimes$  denotes the Hadamard matrix product operation, and the weighted output  $o$  which shape is  $\{W \times H, C\}$  is reshaped to the same size of inputs. In addition, the weighted output is multiplied by a learnable scale parameter and added back the input feature map. The final result of attention module is as following:

$$y = \gamma o + x \quad (4)$$

where  $\gamma$  is initialized as 0. With gradually learning, it will learn to assign more weights to attention maps

### 3.4 Recurrent convolutional layer

Recurrent convolutional layer, which is proposed in Ref.[19], is an important module of our network structure. As shown in Fig. 3, recurrent connections in RCL are utilized to reuse input feature maps for more local semantic information. With the

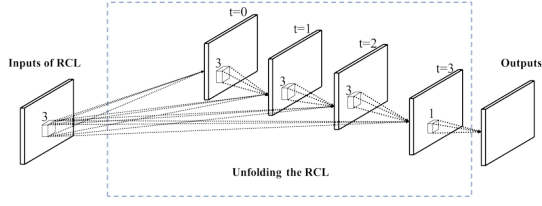


Fig. 3 The structure of RCL.

change of time steps, states of RCL units will evolve. At location  $(i, j)$  on  $k$ th feature maps, the activity of unit is given by:

$$x_{ijk}(t) = g(f(z_{ijk}(t))) \quad (5)$$

where  $t$  is the time steps and  $z_{ijk}$  is the input of the RCL unit.  $z_{ijk}$  is computed by a recurrent connection and the original feed-forward input.  $z_{ijk}$  can be obtained as follows:

$$z_{ijk} = (w_k^f)^T u^{(i,j)}(t) + (w_k^r)^T x^{(i,j)}(t-1) + b_k \quad (6)$$

In Eq. 6,  $u^{(i,j)}$  represents the feed-forward input from previous layer, and  $x^{(i,j)}(t-1)$  denotes recurrent input at time step  $t-1$ . And  $u^{(i,j)}$  and  $x^{(i,j)}(t-1)$  are both vectorized patches at  $(i, j)$  of the feature maps.  $w_k^f$  and  $w_k^r$  are the corresponding weights to feed-forward input and recurrent input respectively.  $b_k$  is a bias for RCL unit. In Eq. 5  $f$  and  $g$  are activation function and local response normalization (LRN) function [15] respectively.  $f$  is the rectified linear activation function as follows

$$f(z_{ijk}(t)) = \max(z_{ijk}(t), 0) \quad (7)$$

and the LRN function  $g$  is given by:

$$g(f_{ijk}(t)) = \frac{f_{ijk}(t)}{\left(1 + \frac{\alpha}{N} \sum_{k'=\max(0, k-N/2)}^{\min(K, k+N/2)} (f_{ijk'})^2\right)^\beta} \quad (8)$$

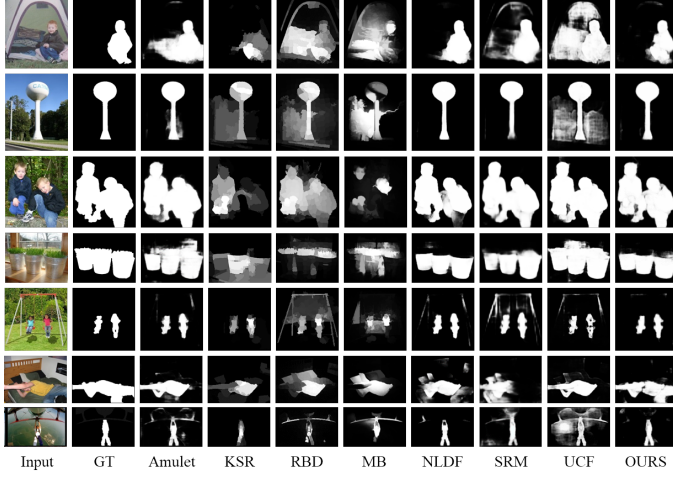
where  $f_{ijk}(t)$  is a abbreviate representation as  $f(z_{ijk}(t))$

## 4 Experiment

### 4.1 Datasets

We conduct performance evaluations on four widely used datasets for saliency detection. ECSSD [23] dataset has 1000 images in total. Most of ECSSD dataset is nature images which are structurally complex. PASCAL-S [18] dataset consists of 850 images selected from PASCAL VOC 2010 dataset without modification. This dataset generally include different semantic objects which is a challenge for saliency detection. DUT-OMRON [38] includes 5168 challenging images, each of which contains one or more salient objects with a complex background. HKU-IS [17] is a large dataset which contains 4447 challenging images, most of which are under low contrast.





**Fig. 4** A visual comparison of our to the other 7 methods.

#### 4.2 Evaluation metrics

We use three different metrics to evaluate our models. Precision-recall(PR) curve is used to evaluate the performance of different methods in terms of precision and recall rate. These two rates could be obtained by calculating the correct classified pixels proportion in the ground truths and detection results respectively. A threshold could affect the calculation of PR value. Thus the PR curve could be plotted with the change of the threshold. Moreover, F-measure score is adopted to comprehensively consider precision and recall rates. By given a fixed threshold, the corresponding *Precision* and *Recall* could be obtained, thus F-measure is given by:

$$F_{\beta} = \frac{(1 + \beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision + Recall} \quad (9)$$

where  $\beta^2$  is set to 0.3 as suggested by previous works [2] for stressing the importance of the precision value. In addition, mean absolute error (MAE), which is the average pixel-wise absolute difference between the saliency map and the binary ground truth, is also utilized to evaluate different models. The MAE score can be computed by:

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

where  $S$  is the saliency map,  $G$  is the binary ground truth,  $W$  and  $H$  are width and height of saliency map  $S$ .

#### 4.3 Implementation details

The proposed algorithm in this paper was implemented in Tensorflow[1]. The weights of our backbone VGG16 network were pre-trained on ImageNet [9], and weights of

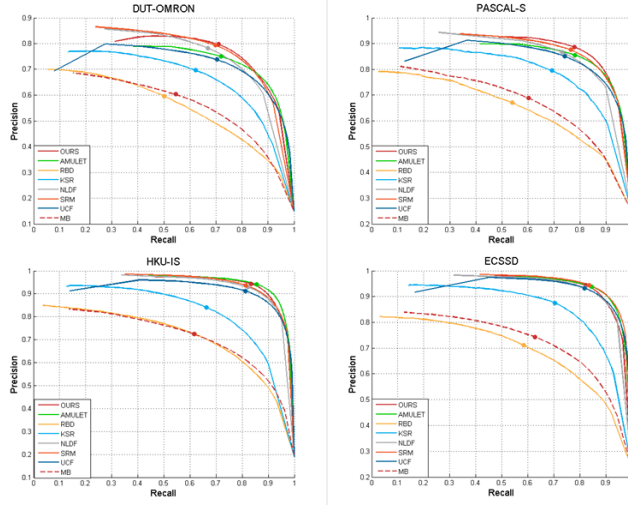


Fig. 5 Precision recall curves on four widely used datasets.

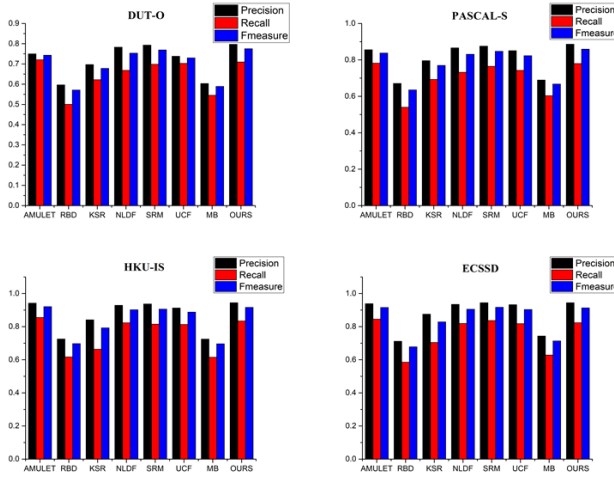
other newly added layers were initialized using the methods introduced in Ref. [10], the biases were initialized to 0. We used Adam optimizer to train our model with the following parameters: initial learning rate of 0.0001,  $\beta_1 = 0.9$ , and  $\beta_2 = 0.999$ . When fed into our model, each image was resized to  $224 \times 224$ , and subtracted a mean pixel value of VGG16.

When training our methods, we followed the training setup of Ref.[20]. 6000 images which were randomly selected from MSRA10K dataset [7] were combined with 3500 randomly selected images from DUT-OMRON dataset to be used as the training set. The training set was horizontally flipped as data augmentation. The rest images and other datasets were used for model test. The training and testing processing were both conducted on a computer with Intel i7-7700k and 32G RAM. An NVIDIA TITAN XP GPU was used to accelerate both training and testing.

#### 4.4 Comparison with State-of-the-arts

Our method is compared with 7 existing state-of-the-art saliency detection methods, including Amulet [44], RBD [25], KSR [27], NLDF [21], SRM [26], UCF[43] and MB [42]. Most saliency maps of them were provided by the authors, and few are implemented by us using the recommended settings.

*Visual Comparison* We provide a visual comparison of our method with the aforementioned approaches in Fig 4. It can be observed that the salient maps generated by our method are subtler than the other methods, and most of our results are very close to the ground truth. It is also worth mentioning that the self-attention give salient regions more weights, which play an important role in our network structure, could efficiently help out model to locate salient regions. And the recurrent structure could effectively enhance local information to detect more details of salient regions.



**Fig. 6** The maximum F-measure, corresponding precision and recall of different methods

**Table 1** MAE(lower is better) and F-measure(higher is better) comparisons with 7 methods on four open datasets. The top three results are highlighted in red, green, and blue fonts respectively.

	DUT-OMRON		PASCAL-S		HKU-IS		ECSSD	
	MAE	F-measure	MAE	F-measure	MAE	F-measure	MAE	F-measure
Amulet	0.0976	0.7429	0.0997	0.8373	0.0507	0.9204	0.0589	0.9154
RBD	0.1609	0.5712	0.2030	0.6353	0.1484	0.6970	0.1769	0.6778
KSR	0.1306	0.6781	0.1540	0.7694	0.1201	0.7922	0.1322	0.8287
UCF	0.1203	0.7296	0.1155	0.8230	0.0620	0.8877	0.0691	0.9034
MB	0.1566	0.5887	0.1955	0.6673	0.1482	0.6961	0.1741	0.7133
NLDF	0.0796	0.7532	0.0977	0.8309	0.0477	0.9020	0.0626	0.9050
SRM	0.0694	0.7690	0.0835	0.8473	0.0459	0.9058	0.0544	0.9172
Ours	0.0714	0.7750	0.0815	0.8591	0.0451	0.9163	0.0637	0.9133

**PR curve** We compare our method with the existing methods in terms of PR curve. As shown in Fig. 5, our method has a best performance on most datasets. On the HKU-IS dataset, the curve of our methods is very close to Amulet, which is the top one on this dataset. The PR curves illustrate that our method is more accurate and reliable, which is reflected in that our method has the highest precision rate on the four datasets.

**F-measure and MAE** We also calculate F-measure and MAE of our method and other existing methods. The F-measure with corresponding precision and recall rate is shown in Fig 6. The F-measure value of our method is the highest on two datasets, and top 3 on the other two datasets. And MAE of our method is also the best on two datasets. The details of F-measure and MAE are shown in Table.1. It can be observed that our model have a good performance in terms of F-measure and MAE.

## 5 Conclusions

In this paper, we proposed a novel self-attention recurrent network for saliency detection. Self-attention module of the network could effectively enhance the global information of deep layers, and the recurrent convolutional structure could improve the availability of shallow layers. Experimental results demonstrate the effectiveness of our network.

## References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: A system for large-scale machine learning. In: Proceedings of the 12th USENIX Conference on Operating Systems Design and Implementation, OSDI16, p. 265283. USENIX Association (2016). URL <http://dl.acm.org/citation.cfm?id=3026877.3026899>
2. Achantay, R., Hemamiz, S., Estraday, F., Süsstrunk, S.: Frequency-tuned salient region detection. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009, pp. 1597–1604 (2009). DOI 10.1109/CVPRW.2009.5206596
3. Bi, S., Li, G., Yu, Y.: Person re-identification using multiple experts with random subspaces. International Journal of Image and Graphics 2(2), 151–157 (2014)
4. Borji, A., Frintrop, S., Sihite, D.N., Itti, L.: Adaptive object tracking by learning background context. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, pp. 23–30 (2012). DOI 10.1109/CVPRW.2012.6239191
5. Cheng, M., Zhang, F., Mitra, N., Huang, X., Hu, S.: RepFinder: Finding Approximately Repeated Scene Elements for Image Editing. ACM Transactions on Graphics TOG 29(4), 1 (2010). DOI 10.1145/1778765.1778820. URL <http://discovery.ucl.ac.uk/1327991/>
6. Cheng, M.M., Hou, Q.B., Zhang, S.H., Rosin, P.L.: Intelligent visual media processing: when graphics meets vision. Journal of Computer Science and Technology 32(1), 110–121 (2017)
7. Cheng, M.M., Zhang, G.X., Mitra, N.J., Huang, X., Hu, S.M.: Global Contrast based Salient Region Detection. pp. 409–416 (2011). DOI 10.1109/CVPR.2011.5995344
8. Chenlei Guo, Liming Zhang: A Novel Multiresolution Spatiotemporal Saliency Detection Model and Its Applications in Image and Video Compression. IEEE Transactions on Image Processing 19(1), 185–198 (2010). DOI 10.1109/TIP.2009.2030969. URL <http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.htm?arnumber=5223506>
9. Deng, J., Dong, W., Socher, R., Li, L., Li, K., Fei-Fei, L.: Imagenet: A large-scale hierarchical image database. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, pp. 248–255 (2009). DOI 10.1109/CVPR.2009.5206848
10. Glorot, X., Bengio, Y.: Understanding the difficulty of training deep feedforward neural networks. In: Y.W. Teh, M. Titterton (eds.) Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics, *Proceedings of Machine Learning Research*, vol. 9, pp. 249–256. PMLR, Chia Laguna Resort, Sardinia, Italy (2010). URL <http://proceedings.mlr.press/v9/glorot10a.html>
11. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.: Deeply Supervised Salient Object Detection with Short Connections (2018). DOI 10.1109/TPAMI.2018.2815688
12. Hua, Y., Zhao, Z., Tian, H., Guo, X., Cai, A.: A probabilistic saliency model with memory-guided top-down cues for free-viewing. In: IEEE International Conference on Multimedia and Expo, pp. 1–6 (2013)
13. Itti, L., Koch, C., Niebur, E.: A Model of Saliency Based Visual Attention for Rapid Scene Analysis. IEEE Transactions on Pattern Analysis and Machine Intelligence 20(11), 1254–1259 (1998). DOI 10.1016/S1053-5357(00)00088-3
14. Koch, C., Ullman, S.: Shifts in selective visual attention: towards the underlying neural circuitry. Human neurobiology 4(4), 219–27 (1985). DOI 10.1016/j.imavis.2008.02.004. URL <http://www.ncbi.nlm.nih.gov/pubmed/3836989>
15. Krizhevsky, A., Sutskever, I., Hinton, G.E.: ImageNet Classification with Deep Convolutional Neural Networks. Advances In Neural Information Processing Systems pp. 1–9 (2012). DOI <http://dx.doi.org/10.1016/j.procy.2014.09.007>

16. Kuen, J., Wang, Z., Wang, G.: Recurrent attentional networks for saliency detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), p. 36683677 (2016). DOI 10.1109/CVPR.2016.399
17. Li, G., Yu, Y.: Deep contrast learning for salient object detection. In: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 2016-January, pp. 478 – 487. Las Vegas, NV, United states (2016)
18. Li, Y., Hou, X., Koch, C., Rehg, J., Yuille, a.: The secrets of salient object segmentation. pp. 4321–4328 (2014). DOI 10.1109/CVPR.2014.43. URL [http://www.stat.ucla.edu/~yuille/Pubs10\\_{\\_}12/LiHouKochRehgYuille.pdf](http://www.stat.ucla.edu/~yuille/Pubs10_{_}12/LiHouKochRehgYuille.pdf)
19. Liang, M., Hu, X.: Recurrent convolutional neural network for object recognition. pp. 3367–3375. IEEE Computer Society (2015). DOI 10.1109/CVPR.2015.7298958. URL <https://arxiv.org/abs/1704.07709>
20. Liu, N., Han, J.: DHSNet: Deep Hierarchical Saliency Network for Salient Object Detection. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 678–686 (2016). DOI 10.1109/CVPR.2016.80. URL <http://ieeexplore.ieee.org/document/7780449/>
21. Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S., Jodoin, P.: Non-local deep features for salient object detection. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 6593–6601 (2017). DOI 10.1109/CVPR.2017.698
22. Ma, Y.F., Lu, L., Zhang, H.J., Li, M.: A user attention model for video summarization. In: Proceedings of the Tenth ACM International Conference on Multimedia, MULTIMEDIA '02, pp. 533–542. ACM, New York, NY, USA (2002). DOI 10.1145/641007.641116. URL <http://doi.acm.org/10.1145/641007.641116>
23. Simonyan, K., Zisserman, A.: Very Deep Convolutional Networks for Large-Scale Image Recognition. pp. 1–14 (2015). DOI 10.1016/j.infsof.2008.09.005. URL <http://arxiv.org/abs/1409.1556>
24. Wang, L., Wang, L., Lu, H., Zhang, P., Xiang, R.: Saliency detection with recurrent fully convolutional networks. In: European Conference on Computer Vision, pp. 825–841 (2016)
25. Wang, Q., Zheng, W., Piramuthu, R.: GraB: Visual Saliency via Novel Graph Model and Background Priors. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 535–543 (2016). DOI 10.1109/CVPR.2016.64. URL <http://ieeexplore.ieee.org/document/7780433/>
26. Wang, T., Borji, A., Zhang, L., Zhang, P., Lu, H.: A stagewise refinement model for detecting salient objects in images. In: 2017 IEEE International Conference on Computer Vision (ICCV), pp. 4039–4048 (2017). DOI 10.1109/ICCV.2017.433
27. Wang, T., Zhang, L., Lu, H., Sun, C., Qi, J.: Kernelized subspace ranking for saliency detection. In: B. Leibe, J. Matas, N. Sebe, M. Welling (eds.) Computer Vision – ECCV 2016, pp. 450–466. Springer International Publishing, Cham (2016)
28. Wang, Y., Lin, X., Wu, L., Zhang, W.: Effective multi-query expansions: Collaborative deep networks for robust landmark retrieval. IEEE Transactions on Image Processing **26**(3), 1393–1404 (2017). DOI 10.1109/TIP.2017.2655449
29. Wang, Y., Lin, X., Wu, L., Zhang, W., Zhang, Q., Huang, X.: Robust subspace clustering for multi-view data by exploiting correlation consensus. IEEE Transactions on Image Processing **24**(11), 3939–3949 (2015). DOI 10.1109/TIP.2015.2457339
30. Wang, Y., Wu, L.: Beyond low-rank representations: Orthogonal clustering basis reconstruction with optimized graph structure for multi-view spectral clustering. Neural Networks **103**, 1 – 8 (2018). DOI <https://doi.org/10.1016/j.neunet.2018.03.006>. URL <http://www.sciencedirect.com/science/article/pii/S0893608018300911>
31. Wang, Y., Wu, L., Lin, X., Gao, J.: Multiview spectral clustering via structured low-rank matrix factorization. IEEE Transactions on Neural Networks and Learning Systems pp. 1–11 (2018). DOI 10.1109/TNNLS.2017.2777489
32. Wang, Y., Zhang, W., Wu, L., Lin, X., Fang, M., Pan, S.: Iterative views agreement: An iterative low-rank based structured optimization method to multi-view spectral clustering. In: Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI'16, pp. 2153–2159. AAAI Press (2016). URL <http://dl.acm.org/citation.cfm?id=3060832.3060922>
33. Wang, Y., Zhang, W., Wu, L., Lin, X., Zhao, X.: Unsupervised metric fusion over multiview data by graph random walk-based cross-view diffusion. IEEE Transactions on Neural Networks and Learning Systems **28**(1), 57–70 (2017). DOI 10.1109/TNNLS.2015.2498149
34. Wang, Y., Zhao, Q.: Superpixel tracking via graph-based semi-supervised svm and supervised saliency detection. In: IEEE International Conference on Multimedia and Expo, pp. 1–6 (2015)

35. Wu, L., Wang, Y., Gao, J., Li, X.: Deep adaptive feature embedding with local sample distributions for person re-identification. *Pattern Recognition* **73**, 275–288 (2018)
36. Wu, L., Wang, Y., Li, X., Gao, J.: Deep attention-based spatially recursive networks for fine-grained visual recognition. *IEEE Transactions on Cybernetics* pp. 1–12 (2018). DOI 10.1109/TCYB.2018.2813971
37. Wu, L., Wang, Y., Li, X., Gao, J.: What-and-where to match: Deep spatially multiplicative integration networks for person re-identification. *Pattern Recognition* **76**, 727–738 (2018)
38. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pp. 3166–3173 (2013). DOI 10.1109/CVPR.2013.407
39. Yang, J.: Top-down visual saliency via joint crf and dictionary learning. In: *Computer Vision and Pattern Recognition*, pp. 2296–2303 (2012)
40. Zhang, G.X., Cheng, M.M., Hu, S.M., Martin, R.R.: A shape-preserving approach to image resizing. *Computer Graphics Forum* **28**(7), 1897–1906 (2009). DOI 10.1111/j.1467-8659.2009.01568.x
41. Zhang, H., Goodfellow, I., Metaxas, D., Odena, A.: Self-Attention Generative Adversarial Networks (2018). URL <http://arxiv.org/abs/1805.08318>
42. Zhang, J., Sclaroff, S., Lin, Z., Shen, X., Price, B., Mech, R.: Minimum barrier salient object detection at 80 FPS. pp. 1404–1412 (2016). DOI 10.1109/ICCV.2015.165
43. Zhang, P., Wang, D., Lu, H., Wang, H., Yin, B.: Learning uncertain convolutional features for accurate saliency detection. In: *2017 IEEE International Conference on Computer Vision (ICCV)*, pp. 212–221 (2017). DOI 10.1109/ICCV.2017.32
44. Zhang, P., Wang, L., Wang, D., Lu, H., Shen, C.: Agile Amulet: Real-Time Salient Object Detection with Contextual Attention (2018). URL <http://arxiv.org/abs/1802.06960>
45. Zhang, X., Wang, T., Qi, J., Lu, H., Wang, G.: Progressive Attention Guided Recurrent Network for Salient Object Detection. In: *Cvpr*, pp. 714–722 (2018). DOI 10.1109/CVPR.2018.00081. URL <https://github.com/zhangxiaoning666/PAGR>
46. Zhu, L., Klein, D.A., Frintrop, S., Cao, Z., Cremers, A.B.: A multisize superpixel approach for salient object detection based on multivariate normal distribution estimation. *IEEE Transactions on Image Processing* **23**(12), 5094–5107 (2014). DOI 10.1109/TIP.2014.2361024