Machine Vision and Applications (2021) 32:45
https://doi.org/10.1007/s00138-021-01172-y

ORIGINAL PAPER

# 3MNet: Multi-task, multi-level and multi-channel feature aggregation network for salient object detection

Xinghe Yan[1] · Zhenxue Chen[1,2] · Q. M. Jonathan Wu[3] · Mengxu Lu[1] · Luna Sun[1]

Received: 9 July 2020 / Revised: 2 January 2021 / Accepted: 22 January 2021 / Published online: 18 February 2021

**Abstract**

Salient object detection is a hot spot of current computer vision. The emergence of the convolutional neural network (CNN) greatly improves the existing detection methods. In this paper, we present 3MNet, which is based on the CNN, to make the utmost of various features of the image and utilize the contour detection task of the salient object to explicitly model the features of multi-level structures, multiple tasks and multiple channels, so as to obtain the final saliency map of the fusion of these features. Specifically, we first utilize contour detection task for auxiliary detection and then utilize use multi-layer network structure to extract multi-scale image information. Finally, we introduce a unique module into the network to model the channel information of the image. Our network has produced good results on five widely used datasets. In addition, we also conducted a series of ablation experiments to verify the effectiveness of some components in the network.

**Keywords** Salient object detection · Fusion model · Multi-level · Multi-task · Multi-channel · Deep neural network · Contour detection

## 1 Introduction

Salient object detection refers to the separation of objects that can most attract human visual attention from background images [1]. Recently, due to the rapid increase in the quantity and quality of image files, salient object detection has become increasingly important as a precondition of various image processing approaches. In the early stage, salient object detection was applied to image content editing [2], object recognition [3], image classification [43] and semantic segmentation [4]. In recent years, it has also played an important role in intelligent photography [5] and image retrieval [6]. It is worth noting that we have seen an interesting application of saliency detection in the emerging Internet video technology. Video site users especially the young like to post their own comments while watching the video. These comments will be displayed on the screen. We call this "bullet screen." In addition, salient object detection is also applied to virtual background technology, which can protect the privacy of users in video conferences, especially during the epidemic of COVID-19. As shown in Fig. 1, our saliency detection technology can help us highlight the important people or objects in the scene so that they are not obscured by the bullet screen, and the real background in video conferencing has been replaced by a virtual background.

Early saliency detection techniques were mainly based on the extraction of certain artificial features. Limited by prior knowledge, these methods sometimes cannot achieve better results in natural scenes. We focus on making full use of deep information at different levels and modeling the image with multi-level mine the information.

Xinghe Yan and Zhenxue Chen have contributed equally.

✉ Zhenxue Chen
chenzhenxue@sdu.edu.cn

Xinghe Yan
201834581@mail.sdu.edu.cn

Q. M. Jonathan Wu
jwu@uwindsor.ca

Mengxu Lu
201500171046@mail.sdu.edu.cn

Luna Sun
201934507@mail.sdu.edu.cn

[1] School of Control Science and Engineering, Shandong University, Jinan 250061, China

[2] Shenzhen Research Institute of Shandong University, Shandong University, Shenzhen 518057, China

[3] Department of Electrical and Computer Engineering, University of Windsor, Windsor N9B 3P4, Canada

**Fig. 1** Applications of salient object detection in video technology. These images are selected from the Chinese video site “Bilibili” and the video conferencing software “Zoom”

Convolutional neural networks can effectively extract the features of the image. The low-level layers usually have smaller receptive fields and can focus on local details of the image, such as edge information. However, unlike the edge detection in traditional tasks, we mainly focus on salient objects and ignore the cluttered lines in the background; as such, we use salient foreground contours as an auxiliary task for our salient object detection.

Most existing methods simply merge multi-channel feature maps, ignoring the variety of effects that different feature channels may have on the final saliency map. We model the feature channels explicitly, introduce a global pooling method with a large visual receptive field into the modeling of the feature channels and reweight each feature channel.

In general, our proposed 3MNet uses a U-shaped structure as the main structure, with contour detection branches as auxiliary tasks, and introduces channel reweighting modules in the network structure, so as to explicitly model and combine the multi-task, multi-level and multi-channel features of the image. Specifically, the contour detection task can refine the edge details of salient objects. The multi-level network structure can better aggregate the local and global feature information of the image. Multiple multi-channel feature maps are generated in the deep network. Modeling the channel features helps to mine the deep channel information in the image and enhance the weight of high contribution channels. Our subsequent experiments also proved that combining multiple image features can effectively improve the detection accuracy.

The main contributions of this paper are as follows:

(1) The proposed 3MNet makes full use of the deep salient information in the image and combines the multi-task, multi-level and multi-channel features to explicitly model the saliency detection task. We have achieved good results on the basis of salient object detection tasks, supplemented by target contour detection.

(2) Compared with traditional models and some other deep detection models, our model has higher accuracy, and multiple evaluation indicators on the five most commonly used data sets are ahead of other methods. In addition, we conducted a series of ablation experiments to verify the effectiveness of our network structure.

(3) Our training process requires saliency object contour information. Therefore, we provide saliency target contour ground-truth maps of multiple training sets as a supplement to the training set, so that researchers can adopt more optional auxiliary methods for saliency detection.

The rest part of our paper is organized as follows: Section 2 introduces the related works of salient object detection. The specific structure of our proposed approach are described in Sect. 3. Section 4 shows and analyzes the results of our experiment. Section 5 makes a conclusion to our paper.

## 2 Related works

Early salient object detection used a data-driven bottom-up approach. In 1998, Itti et al. [7] proposed the classic saliency visual attention model. For a long time, manual features such as contrast, color and background prior dominated the salient object detection.

Achanta et al. [8] introduced a frequency-tuned model to extract the global features of the image. Jiang et al. [9] used the absorbing Markov chain to calculate the absorption time. They considered solving problems mathematically rather than imitating human vision. [42] introduced a bootstrap learning algorithm into salient object detection task. Researchers also proposed methods of preprocessing and post-processing such as the super-pixel detection [10] and the conditional random field [11] methods.

Recently, salient object detection models based on deep learning have been widely studied and applied. Inspired by various network optimization methods, especially the emergence of convolutional neural network structures [24], more and more models designed for saliency detection tasks are appearing and have achieved unprecedented detection

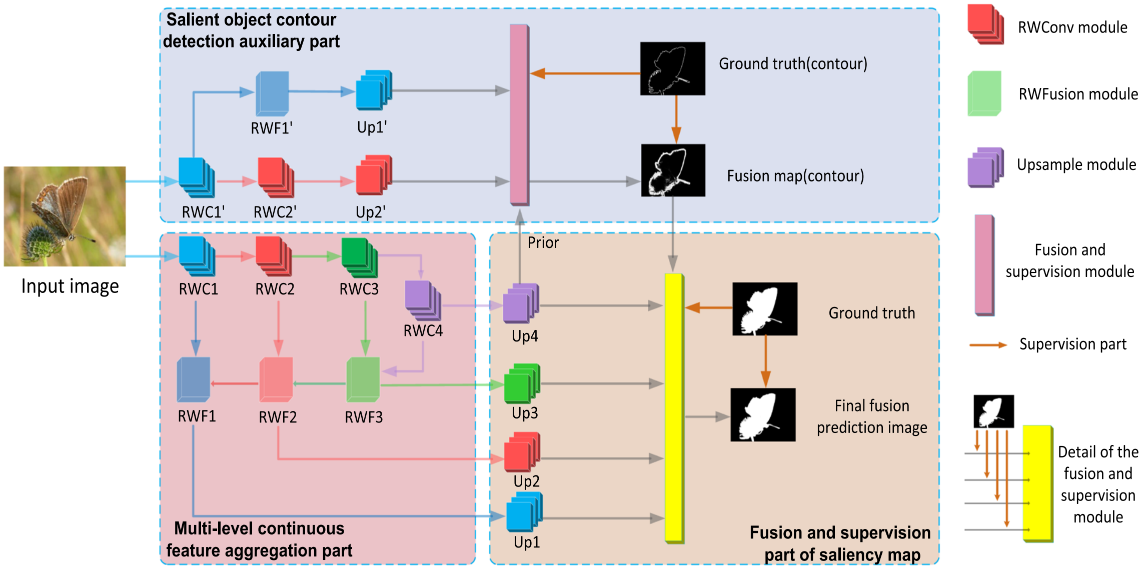

Salient object contour detection auxiliary part
RWF1'
Up1'
RWC1'
RWC2'
Up2'
Ground truth(contour)
Fusion map(contour)
Input image
Prior
RWC1
RWC2
RWC3
RWC4
RWF1
RWF2
RWF3
Up4
Up3
Up2
Up1
Ground truth
Final fusion prediction image
Multi-level continuous feature aggregation part
Fusion and supervision part of saliency map
RWConv module
RWFusion module
Upsample module
Fusion and supervision module
Supervision part
Detail of the fusion and supervision module

**Fig. 2** Overall structure of our proposed network framework. The RWC module is the RWConv module. The upper part explicitly models the contour information and uses this information to help detect salient targets. The lower left part uses image multi-level features to fuse salient feature maps

effects on various evaluation criteria. Since the introduction of VGG [12] and residual networks [13], saliency detection models with these networks as the base structure have developed considerably. Researchers have achieved better results by appropriately increasing the depth of the network and expanding the width of the network. [14] combine features of different levels in the deep network to predict salient regions. DHSNet [15] aggregates the characteristics of many different receptive fields to obtain performance gains. Ronneberger et al. [16] propose a U-shaped network structure for image segmentation. Liu et al. propose PoolNet [17] for saliency detection based on a similar structure and obtained accurate and fast detection performance. Hou et al. [18] ingeniously build short connections between multi-level feature maps to make full use of high-level features to guide detection. Li et al. [41] explore the channel characteristics with reference to the structure of SENet [23].

Apart from innovations in depth and breadth in the network structure, some researchers have also attempted multi-task-assisted saliency detection. Li et al. [19] combine the saliency detection task with the image semantic segmentation task. Through the collaborative feature learning of these two related tasks, the shared convolutional layer produces effective object perception features. Zhuge et al. [20] focus on using the boundary features of the objects in the image, utilizing edge truth labels to supervise and refine the details of the detection feature map. [44] make full use of the multi-temporal features and show the effectiveness of multiple features in improving detection performance. [21] apply saliency detection to dynamic video processing, greatly expanding the application space of saliency detection.

## 3 Proposed Approach

Our model captures the features of the image to be detected from the following aspects: First, we set up two sets of network frameworks to perform saliency target detection and salient object contour detection in parallel. Second, we use a U-shaped network construction [16] for the main structure of each network to aggregate the salient features extracted from different levels. Finally, for the basic unit of each convolution module, we make full use of the channel characteristics, use global pooling to obtain the corresponding global receptive field of each channel and learn how much each channel contributes to the salient features. According to the learning results, we then recalibrate the weights of the feature channels. The specific framework of the model is shown in Fig. 2

### 3.1 Multi-channel characteristic response reweighting module

For common RGB three-channel images, each channel's salient stimulation of the human eyes of each channel may be different [22]. This reminds us that different feature channels of salient feature maps may also contribute differently to the saliency detection. We refer to the structure of SENet [23] and propose a similar multi-channel reweighted convolution module RWConv and a multi-channel reweighted fusion module RWFusion. These two structures are shown in Fig. 3.

For each basic convolution unit RWConv, we use ResNet's convolution layer [13] as its main structure. On this basis,

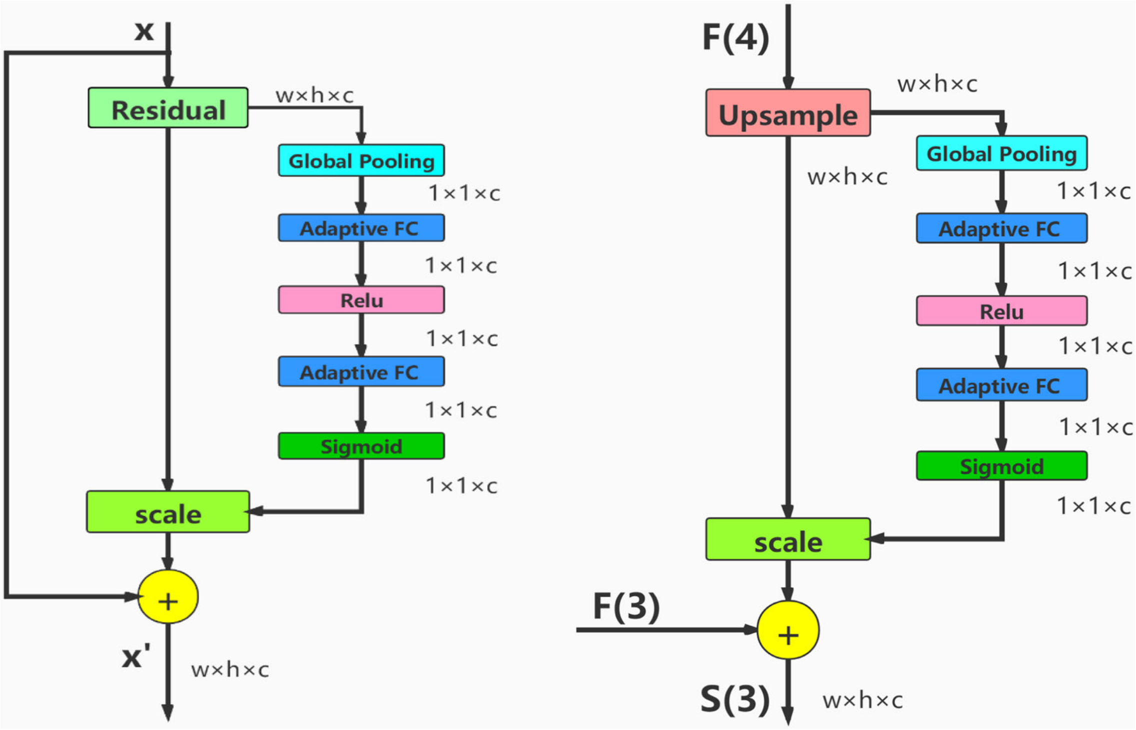


**Fig. 3** Specific structure of the RWConv module (left) and the RWFusion module (right)

we introduce a second branch between the residual and the accumulated sum $x'$, as the weight storage area. For an input image with number of channels c, width w and height h, first, we use global pooling to convert the input to an output of $1 \times 1 \times$c. To some extent, these c real numbers can describe the global characteristics of the input. Its calculation method is shown in Eq. 1.

$$W_k = \frac{1}{w \times h} \sum_{i=1}^{w} \sum_{j=1}^{h} P_k(i, j), \quad k = 1, 2, \ldots, c, \tag{1}$$

where $P_k(i, j)$ is the feature value corresponding to the coordinate $(i, j)$ in the kth channel of the given feature map.

In order to fully represent the relationship between each channel, so that our model can focus on the channels that contribute more, we add 2 fully connected layers after global pooling. The number of fully connected points in each layer is the same as the number of channels in the upper layer, and a Relu layer is added to ensure the nonlinearity of the model. After obtaining the final channel weight $W_c'$, we weight and accumulate the input value corresponding to the c weight parameters to obtain the final output.

$$S_k = p_k \times W_k', \quad k = 1, 2, \ldots, c, \tag{2}$$

where this operation corresponds to the scale module in the network.

The basic structure of the RWFusion part is roughly the same as that of the RWConv part, except that one of the addends x is replaced by the same size feature map on the other side of the U-shaped network. The input of the main part is obtained by the upsampling operation.

The basic module of the contour detection part is the same as that of the above-mentioned RWConv and RWFusion. This fusion method takes into account the multi-level and multi-channel characteristics, makes full use of the detailed information of the image and enhances the expression ability of the network.

### 3.2 Salient object contour detection auxiliary module

Explicitly modeling contour features is undoubtedly helpful for optimizing the details of salient object. However, the high-level feature maps often have large receptive fields and cannot pay attention to the details of the target. Low-level feature maps can help us optimize the contour details of objects [25]. As such, we take low-level features into consideration. We use a two-layer RWConv structure to extract the contour features of the object in the main part of the network; then, after obtaining the significant contour feature map $E_j$, we use the same fusion method. The calculation method is as follows:

$$\begin{aligned} E_{f2} &= up\left(E_2, 4\right), \\ E_{f1} &= up\left(RWF\left(E_1, E_2\right), 2\right), \end{aligned} \tag{3}$$

where $up(*, \theta)$ means upsampling the feature map, $\theta$ is the upsampling multiple and $RWF$ is the multi-channel feature reweighted fusion operation.

We fuse two saliency contour feature maps according to the following combined strategy:

$$E_{fusion} = Conv\left(Con\left(E_{f1}, E_{f2}\right), \omega_i\right), \tag{4}$$

where $Con$ means that the feature maps are concatenated by channel and $Conv$ means the convolution operation. The parameter $\omega_i$ is trained through the convolutional layer.

In order to effectively obtain the salient contours of salient targets, we imitate the prior knowledge in the traditional method [26] and increase the contour weights of salient regions. At this time, we use the high-level feature map $S_4$ as a prior map to emphasize the importance of the saliency region and get the final fusion contour saliency map $E_f$.

### 3.3 Multi-level continuous feature aggregation module

For the main part of the model framework, we adopt a design that is similar to a U-shaped network structure [16]. The basic unit of the convolution layer is a multi-channel feature response reweighting module (RWConv), which we introduced in detail in Sect. 3.1. First, the input image passes through four consecutive levels of RWConv layers to form four corresponding-level feature maps. The feature fusion module at each level is RWFusion, which we have also introduced in Sect. 3.1. We represent the feature map obtained at each level of the saliency target detection section as $F_i$ , and we fuse them according to Eq. 5:

$$\begin{aligned} S_4 &= up\left(F_4, 16\right), \\ S_3 &= up\left(RWF\left(F_3, F_4\right), 8\right), \quad i = 3, \\ S_i &= up\left(RWF\left(F_i, S_{i+1}\right), 2^i\right), \quad i = 1, 2, \\ S_f &= Conv\left(Con\left(S_1, S_{2,} S_3, S_4\right), \omega_i\right), \end{aligned} \tag{5}$$

where the operations in Eq. 5 are the same as the operations in Eq. 3 and Eq. 4.

The final result $R_f$ of the multi-feature fusion is as:

$$R_f = Conv\left(Con\left(S_f, E_{f,}\right), \omega_i\right) \tag{6}$$

## 4 Experiment and analysis

### 4.1 Implementation details

In the training phase, we use the MSRA10K dataset [27] as our training set. The dataset contains 10,000 high-quality images with salient objects and is labeled at the pixel level. In addition, we randomly selected 5000 images from the DUTS-TR [40] dataset to expand our training set. We do not use validation sets during the training phase. Since our training has salient object contour supervision in addition to the original ground-truth map, we need to expand the dataset. We utilize the Laplacian operator in the OpenCV toolbox to perform edge detection on the targets in the ground-truth map. In this way, we get a 10K group of images with pixel-level object contour annotation. Our implementation is based on the pytorch deep learning framework. The training and testing processes are performed on a desktop with an NVIDIA GeForce RTX 2080Ti (with 11G memory). On our desktop, our model can achieve a relatively fast speed of 16 fps. The initial values of the main parameters of the first half of the U-shaped network are consistent with ResNet [13], and the other parameters are initialized randomly. We use the cross-entropy loss function to calculate the loss between the feature map and the truth map. The calculation method of Softmax function and the cross-entropy loss function is as follows:

$$\begin{aligned} p_i &= \frac{e^{\alpha_i}}{e^{\alpha_1} + e^{\alpha_2}}, \\ L(\omega) &= -\sum_{i=1}^{c} y_i \log p_i, \end{aligned} \tag{7}$$

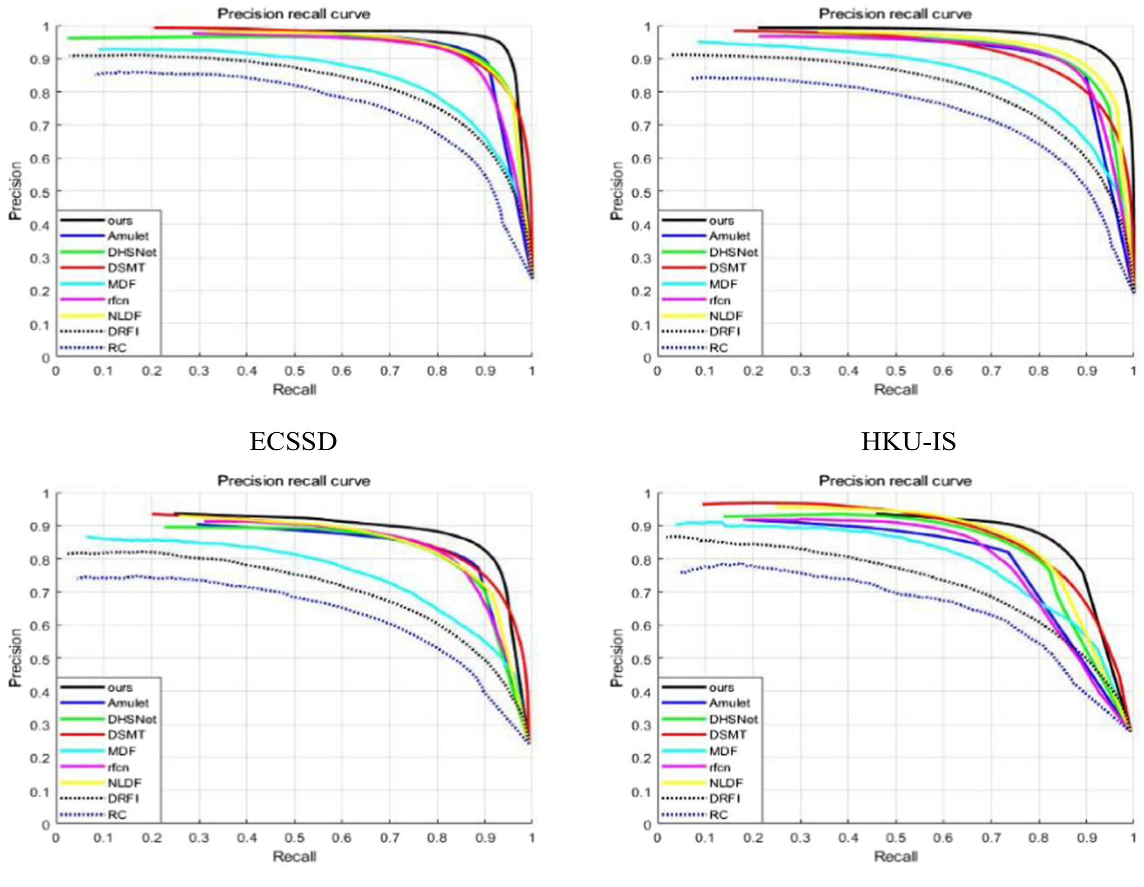
where $\alpha_i$ represents the ith value of the predicted C-dimensional vector and $y_i$ represents the value of the label in the ground truth. We take $C$ as 2 to distinguish background and foreground. $\omega$ is the weight parameter.

The model we propose is end to end and does not contain any preprocessing or post-processing operations. We trained 30 epochs on the network.

During network training, the stochastic gradient descent optimization method is used, the momentum is set to 0.9, and the weight decay is 0.0005. The basic learning rate is set to 1e-6, and it is reduced by 50% every 10 epochs.

### 4.2 Datasets

We qualitatively and quantitatively compare different methods and their performance on five commonly used benchmark datasets. The ECSSD [28] dataset contains 1,000 complex images, and the images contain salient objects of different sizes. The SOD [29] dataset is built on the basis of BSD [30], and pixel-level annotations were made by Jiang et al. [31]. It contains 300 high-quality and challenging images. The DUT-OMRON [32] dataset contains 5,168 high-quality and challenging images. The HKU-IS [33] dataset consists of 4,447 annotated high-quality images, and most of them contain multiple salient objects. The PASCAL-S [34] dataset



ECSSD

HKU-IS

PASCAL-S

SOD

**Fig. 4** P–R curves of some mentioned datasets

contains 850 natural images which are derived from the PASCAL VOC dataset [35].

## 4.3 Evaluation metrics

We use five common evaluation metrics to assess our model performance, including precision–recall curve [1], F-measure [36], receiver operating characteristic curve (ROC) [36], area under ROC curve (AUC) [36] and MAE [36,37]. We binarize the predicted saliency map according to a certain threshold and then compare the obtained binarization map with the ground truth to get the precision and recall, with the F-measure as the harmonic mean of the two. These are calculated as:

$$F_\beta = \left(1+\beta^2\right) \times \frac{Precision \times Recall}{\beta^2 \times Precision + Recall}\ , \tag{8}$$

where $\beta^2$ is generally set to 0.3 in order to emphasize the importance of the precision value [1]. For each fixed binarization threshold, different P–R and F-measure values are obtained. We draw them as curves, and we pick the maximum value of all F-measure calculation results.

Additionally, we can obtain the paired false positive rate ($FPR$) and true positive rate ($TPR$), from which we can get the ROC curve and calculate the AUC value.

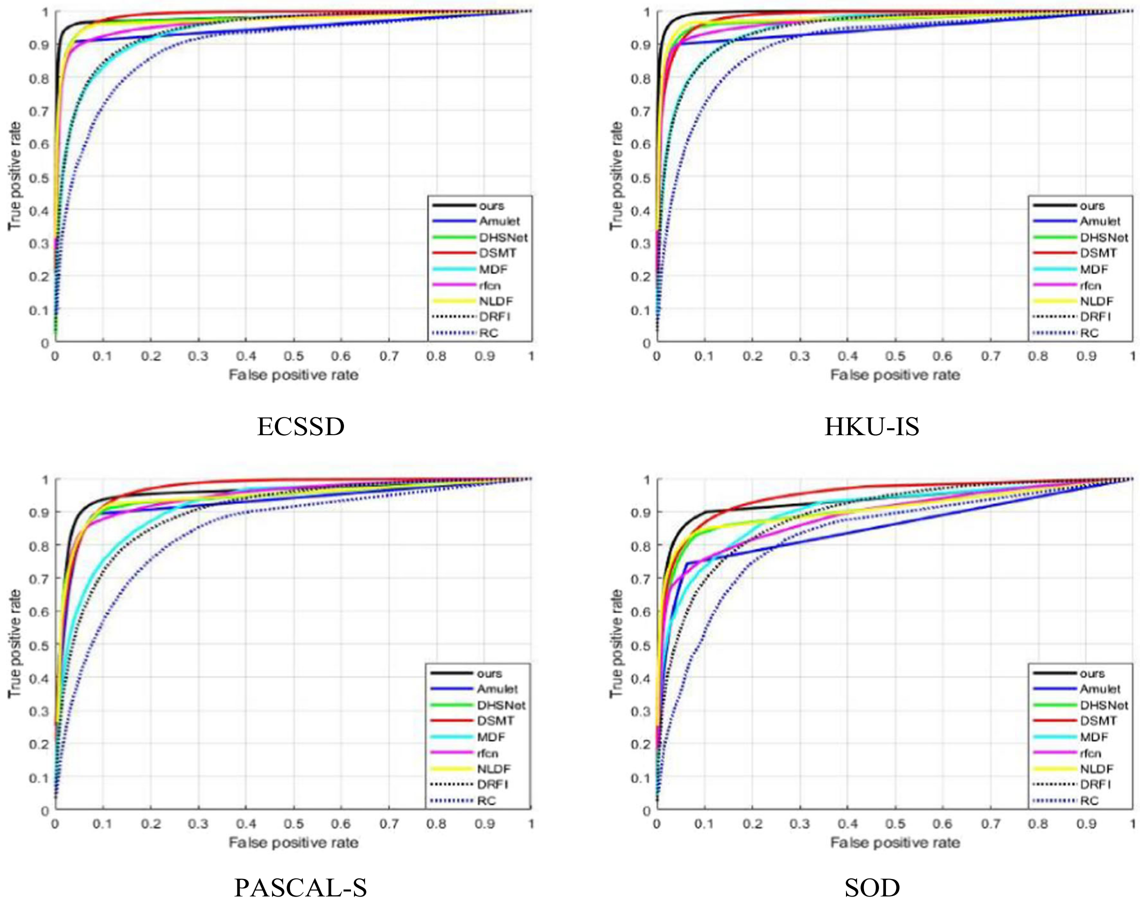
$$TPR = \frac{|M \cap G|}{G},\ FPR = \frac{|M \cap \bar{G}|}{\bar{G}}\ , \tag{9}$$

where $M$ is the binary salient feature map, $G$ is the truth map and $\bar{G}$ is the result of negating $G$.

MAE is expressed as the mean absolute error between the normalized saliency map $S$ and the ground truth $G$. Its calculation formula is as:

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

where $W$ and $H$ are the width and height of the image, respectively.



ECSSD

HKU-IS

PASCAL-S

SOD

**Fig. 5** ROC of four widely used datasets

### 4.4 Comparison with different methods

Our experiments quantitatively compare our model with eight other saliency detection algorithms (Amulet [14], DSMT [19], DHSNet [15], MDF [33], NLDF [38], RFCN [39], DRFI [31] and RC [27]). The P–R curves of some of the mentioned datasets are shown in Fig. 4, and the ROC is shown in Fig. 5. We compare the five performance indicators of the model on the five datasets mentioned above.

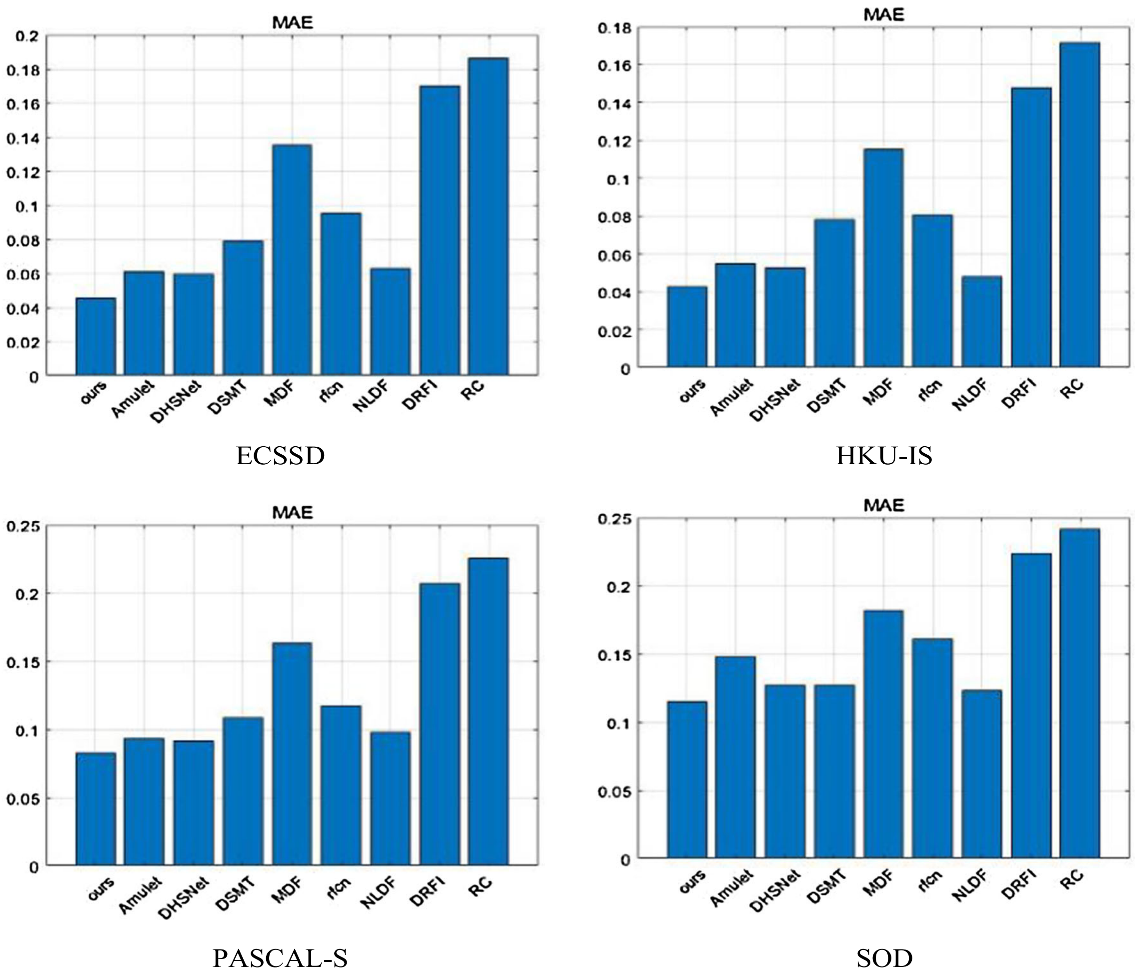
***Quantitative Comparison***: On the five commonly used datasets mentioned above, we quantitatively compare the P–R curve, the ROC curve and the MAE value, and the corresponding experimental results are shown below.

For the P–R curve, the quantitative result we are interested in is the F-measure, and the AUC in the ROC curve can be quantitatively compared as shown in Table 1.

It can be seen from the table that, for the model we proposed, the performance in five popular datasets of its two quantitative indicators' F-measure, AUC, is significantly better than within the other methods. The bold part in the table indicates that the method performs best on the dataset. In particular, the evaluation criteria F-measure, compared with the second place, has an increase of 3.2%, 3.7%, 5.4%, 4.1% and 2.9% on HKU-IS, ECSSD, DUT-OMRON, PASCAL-S and SOD datasets. Although the method DSMT scores higher on the auc indicator on PASCAL-S and SOD datasets, it is not as good as our method in terms of refining the target contour and uniformly highlighting the salient target, which can be found in the following qualitative visual comparison. The DRFI and RC methods are outstanding among the traditional methods. By comparison, we can prove that the models' performance based on the deep network is much better than the traditional method, which is explained in [24].

Figure 6 shows the experimental results of the nine methods we mentioned regarding MAE values in four datasets. And the histogram shows that our model has the best performance on these datasets.

***Qualitative Comparison***: Fig. 7 compares the performance of our model with other detection methods for different scenarios. Our images are selected from the aforementioned datasets. Through intuitive comparison, we can find that, due to the explicit modeling of the contour of the salient object,



ECSSD

HKU-IS

PASCAL-S

SOD

**Fig. 6** MAE histogram of the above detection methods. From left to right in each histogram is our method, Amulet [14], DHSNet [15], DSMT [19], MDF [33], NLDF [38], RFCN [39], DRFI [31] and RC [27]

**Table 1** Quantitative indicators of various advanced detection methods. The best results was bolded

| *Metric* | HKU-IS | | ECSSD | | DUT-OMRON | | PASCAL-S | | SOD | |
|---|---|---|---|---|---|---|---|---|---|---|
| *Methods* | $F_\beta$ | AUC | $F_\beta$ | AUC | $F_\beta$ | AUC | $F_\beta$ | AUC | $F_\beta$ | AUC |
| Ours | **0.931** | **0.985** | **0.939** | **0.977** | **0.815** | **0.955** | **0.862** | 0.957 | **0.861** | 0.933 |
| Amulet [14] | 0.886 | 0.941 | 0.911 | 0.947 | 0.734 | 0.908 | 0.825 | 0.924 | 0.799 | 0.850 |
| DHSNet [15] | 0.891 | 0.968 | 0.907 | 0.973 | \ | \ | 0.821 | 0.938 | 0.823 | 0.904 |
| DSMT [19] | 0.866 | 0.982 | 0.899 | 0.976 | 0.773 | 0.953 | 0.828 | **0.961** | 0.829 | **0.945** |
| MDF [33] | 0.807 | 0.948 | 0.808 | 0.939 | 0.679 | 0.922 | 0.727 | 0.916 | 0.764 | 0.899 |
| NLDF [38] | 0.902 | 0.974 | 0.905 | 0.970 | 0.753 | 0.927 | 0.822 | 0.939 | 0.837 | 0.907 |
| RFCN [39] | 0.888 | 0.968 | 0.898 | 0.966 | 0.747 | 0.918 | 0.827 | 0.942 | 0.805 | 0.886 |
| DRFI [31] | 0.772 | 0.946 | 0.782 | 0.943 | 0.664 | 0.931 | 0.688 | 0.901 | 0.699 | 0.889 |
| RC [27] | 0.718 | 0.898 | 0.738 | 0.893 | 0.601 | 0.859 | 0.640 | 0.842 | 0.657 | 0.823 |

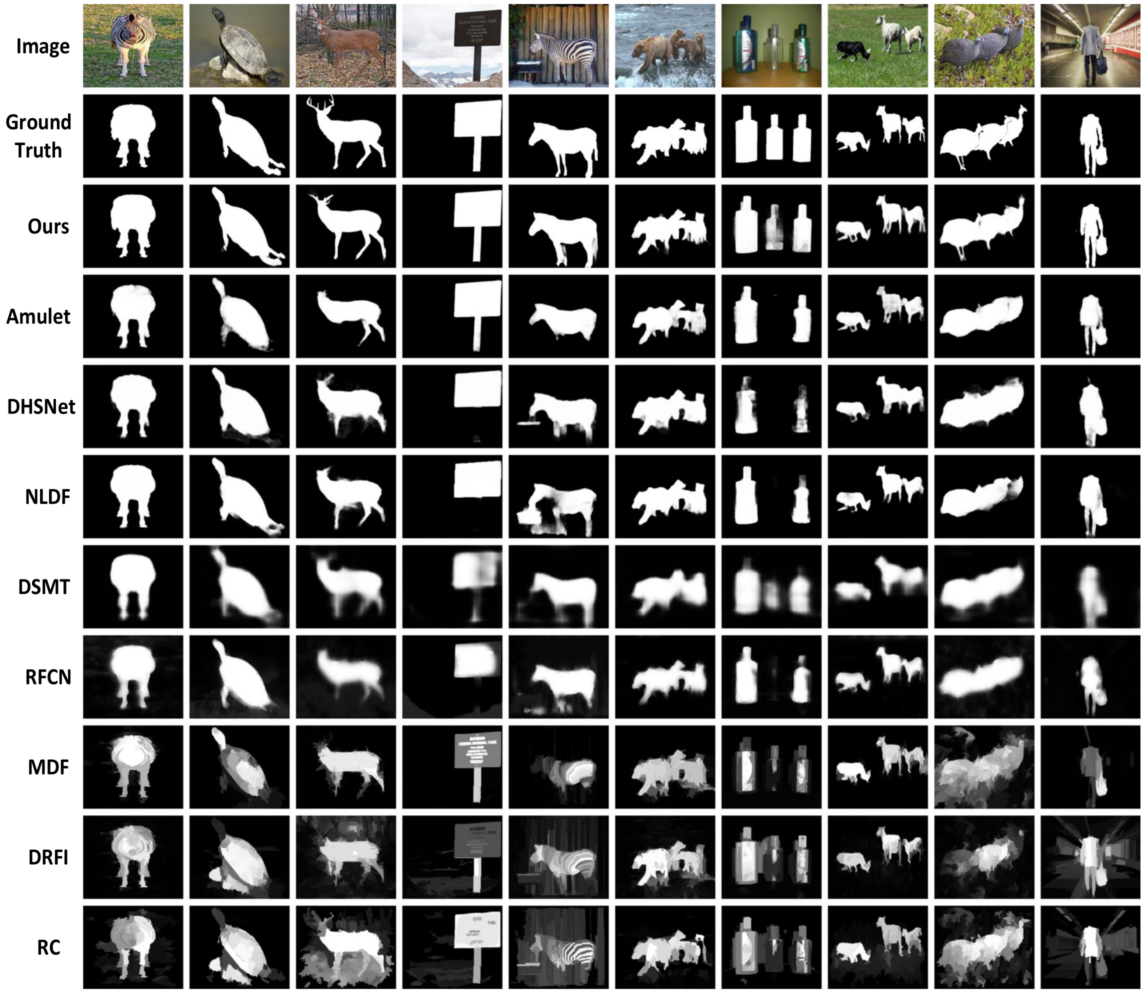


**Fig. 7** Qualitative comparison of our method with other methods in different application scenarios

our method can better refine the contour of the target to be tested; it also achieved good performance in the overall consistency of the salient target.

### 4.5 Ablation experiment

Our ablation experiments focus on the impact of the contour-aided detection and the multi-channel reweighting module on the performance of the detection. Our baseline model is a network structure without these two parts. We take the ECSSD [28] dataset as an example and add contour information to assist detection and channel reweighting modules. The evaluation indicators $F_\beta$ and MAE are shown in Table 2. After successively introducing contour features and channel features, the F-measure has improved by 2.1% and 1.5%, while the MAE has been reduced by 0.012 and 0.002, respectively. From this, we can discover that the contour feature improves the detection performance more significantly.

**Table 2** Changes in quantitative indicators during ablation experiments.(on ECSSD dataset)

| Type | Metric | |
|---|---|---|
| | $F_\beta$ | $MAE$ |
| Base | 0.906 | 0.059 |
| Base+RW | 0.911 | 0.055 |
| Base+Contour | 0.925 | 0.047 |
| Base+RW+Contour | 0.939 | 0.045 |

The salient feature maps before and after the multi-feature cues are added as shown in Figs. 8 and 9. Qualitative observations show that the saliency map with the contour assist module has clearer boundaries. Adding a multi-channel reweighting module can make full use of the information in the feature channels to help highlight the target area uniformly.

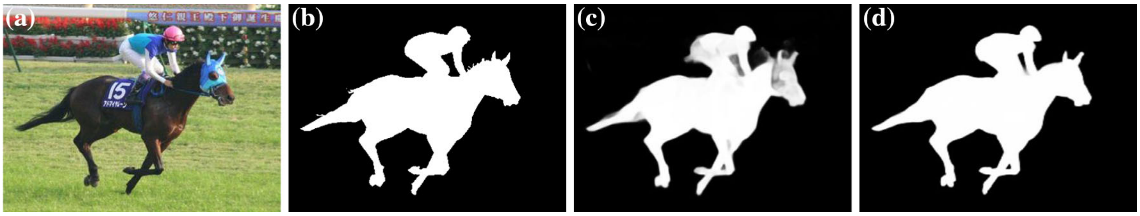


**Fig. 8** Comparison images before and after adding multi-channel features. **a** Input image; **b** ground truth; **c** feature map before adding multi-channel features; **d** feature map with multi-channel features

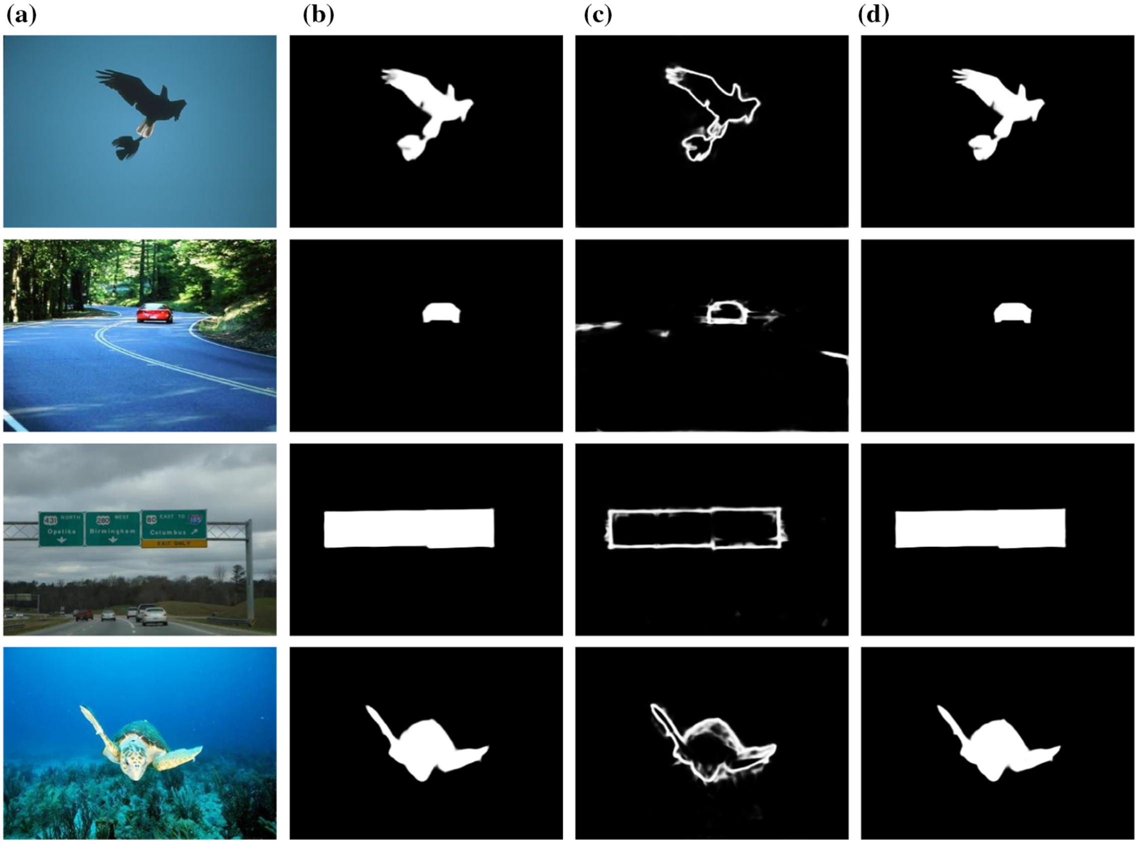


**Fig. 9** Visual effect of adding contour assistant detection module in various of unmanned missions, including aerial photography, intelligent driving, traffic sign detection and underwater target detection. (a) Input images; (b) original detection feature maps; (c) contour auxiliary feature maps; (4) feature maps with contour information. After adding contour information, the detailed information of the object is more refined. For instance, the wings of the bird in the picture become clearer

## 5 Conclusion

This paper explores methods to make full use of multiple aspects of image information and proposes a saliency detection network that combines multi-level, multi-task and multi-channel features. The network explicitly models these three features of the image. Multi-level features are modeled with U-shaped networks, multi-task features are modeled with contour-assisted branches, and multi-channel features are modeled with reweighting modules. The model is an end-to-end model without any preprocessing or post-processing. It is relatively flexible for multi-tasking as well as multi-channel modeling, and it can be used to improve most existing models. Experiments show that our method is comparable to the state-of-the-art deep learning methods on various datasets.

**Acknowledgements** This work was supported in part by the National Natural Science Foundation of China (61876099), in part by the National Key R&D Program of China (2019YFB1311001), in part by the Scientific and Technological Development Project of Shandong Province (2019GSF 111002), in part by the Shenzhen Science and Technology Research and Development Funds (JCYJ20180305164401921), in part by the Foundation of Ministry of Education Key Laboratory of

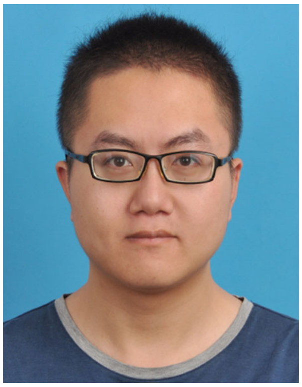
System Control and Information Processing (Scip201801), in part by the Foundation of Key Laboratory of Intelligent Computing & Information Processing of Ministry of Education (2018ICIP03), and in part by the Foundation of State Key Laboratory of Integrated Services Networks (ISN20-06). Xinghe Yan and Zhenxue Chen contributed equally to this work and should be considered as the co-first authors.

## References

1. Borji, A.: What is a salient object? a dataset and a baseline model for salient object detection. IEEE Trans. Image Process. **24**(2), 742–756 (2015)
2. Zhang, G.-X., Cheng, M.-M., Hu, S.-M., Martin, R.R.: A shape-preserving approach to image resizing. Comput. Graphics Forum **28**(7), 1897–1906 (2009)
3. Lu, Y.F., Lim, M.T., Zhang, H.Z., Kang, T.K.: Enhanced hierarchical model of object recognition based on a novel patch selection method in salient regions. Computer Vision Iet **9**(5), 663–672 (2015)
4. Liu, W., Qing, X., Zhou, J.: A novel image segmentation algorithm based on visual saliency detection and integrated feature extraction. In: *International Conference on Communication and Electronics Systems*, pp. 1–5 (2016)
5. Chen, T., Cheng, M.-M., Tan, P., Shamir, A., Hu, S.-M.: Sketch2photo: Internet image montage. *ACM Transactions on Graphics*, vol. 28, no. 5, pp. 124:1–10, (2009)
6. Hussain, C.A., Rao, D.V., Masthani, S.A.: Robust pre-processing technique based on saliency detection for content based image retrieval systems. Proc. Comput. Sci. **85**, 571–580 (2016)
7. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE Trans. Pattern Anal. Mach. Intell. **20**(11), 1254–1259 (1998)
8. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. IEEE Conf. Comput. Vis. Pattern Recogn. **2009**, 1597–1604 (2009)
9. Jiang, B., Zhang, L., Lu, H., Yang, C., Yang, M.: Saliency detection via absorbing markov chain. In *IEEE International Conference on Computer Vision*, pp. 1665–1672 (2013)
10. Liu, Z., Zhang, X., Luo, S., Le Meur, O.: Superpixel-based spatiotemporal saliency detection. IEEE Trans. Circuits Syst. Video Technol. **24**(9), 1522–1540 (2014)
11. Li, G., Yu, Y.: Deep contrast learning for salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 478–487 (2016)
12. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: *International Conference on Learning Representations*, pp. 1–14 (2015)
13. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 770–778 (2016)
14. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In *IEEE International Conference on Computer Vision*, pp. 202–211 (2017)
15. Liu, N., Han, J.: Dhsnet: Deep hierarchical saliency network for salient object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 678–686 (2016)
16. Ronneberger, O., Fischer, P., Brox, T.: U-net: Convolutional networks for biomedical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, pp. 234–241 (2015)
17. Liu, J., Hou, Q., Cheng, M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3912–3921 (2019)
18. Hou, Q., Cheng, M., Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 5300–5309
19. Li, X., Zhao, L., Wei, L., Yang, M., Wu, F., Zhuang, Y., Ling, H., Wang, J.: Deepsaliency: multi-task deep neural network model for salient object detection. IEEE Trans. Image Process. **25**(8), 3919–3930 (2016)
20. Zhuge, Y., Yang, G., Zhang, P., Lu, H.: Boundary-guided feature aggregation network for salient object detection. IEEE Signal Process. Lett. **25**(12), 1800–1804 (2018)
21. Chen, C., Li, S., Wang, Y., Qin, H., Hao, A.: Video saliency detection via spatial-temporal fusion and low-rank coherency diffusion. IEEE Trans. Image Process. **26**(7), 3156–3170 (2017)
22. Yuan, Y., Han, A., Han, F.: Saliency detection based on non-uniform quantification for rgb channels and weights for lab channels. In: *Chinese Conference on Computer Vision*, 2015, pp. 258–266
23. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141
24. Zeiler, M. D., Fergus, R.: Visualizing and understanding convolutional networks. In: *European Conference on Computer Vision*, 2014, pp. 818–833
25. Liu, Y., Cheng, M., Hu, X., Bian, J., Zhang, L., Bai, X., Tang, J.: Richer convolutional features for edge detection. IEEE Trans. Pattern Anal. Mach. Intell. **41**(8), 1939–1946 (2019)
26. Yang, C., Zhang, L., Lu, H.: Graph-regularized saliency detection with convex-hull-based center prior. IEEE Signal Process. Lett. **20**(7), 637–640 (2013)
27. Cheng, M., Zhang, G., Mitra, N. J., Huang, X., Hu, S.: Global contrast based salient region detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2011, pp. 409–416
28. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 1155–1162
29. Movahedi, V., Elder, J. H.: Design and perceptual validation of performance measures for salient object segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2010, pp. 49–56
30. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: *Proceedings Eighth IEEE International Conference on Computer Vision*, vol. 2, 2001, pp. 416–423
31. Jiang, H., Wang, J., Yuan, Z., Wu, Y., Zheng, N., Li, S.: Salient object detection: A discriminative regional feature integration approach. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 2083–2090
32. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.: Saliency detection via graph-based manifold ranking. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2013, pp. 3166–3173
33. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 5455–5463
34. Li, Y., Hou, X., Koch, C., Rehg, J. M., Yuille, A. L.: The secrets of salient object segmentation. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 280–287
35. Everingham, M., Gool, L.V., Williams, C.K.I., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput. Vis. **88**(2), 303–338 (2010)
36. Borji, A., Cheng, M., Jiang, H., Li, J.: Salient object detection: a benchmark. IEEE Trans. Image Process **24**(12), 5706–5722 (2015)

37. Zhu, W., Liang, S., Wei, Y., Sun, J.: Saliency optimization from robust background detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 2814–2821
38. Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S., Jodoin, P.: Non-local deep features for salient object detection. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 6593–6601
39. Wang, L., Wang, L., Lu, H., Zhang, P., Ruan, X.: Saliency detection with recurrent fully convolutional networks. In: *European Conference on Computer Vision*, (2016), pp 825–841
40. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B and Ruan, X.: Learning to detect salient objects with image-level supervision. In: *IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3796–3805
41. Li, C., Chen, Z., Wu, Q.M.J., Liu, C.: Deep saliency with channel-wise hierarchical feature responses for traffic sign detection. IEEE Trans. Intell. Transp. Syst. **20**(7), 2497–2509 (2019)
42. Tong, N., Lu, N., Ruan, X., Yang, M.: Salient object detection via bootstrap learning. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1884–1892 (2015)
43. Sarkar, R., Acton, S.T.: Sdl: Saliency-based dictionary learning framework for image similarity. IEEE Trans. Image Process. **27**(2), 749–763 (2018)
44. Li, X., Shen, H., Zhang, L., Zhang, H., Yuan, Q., Yang, G.: Recovering quantitative remote sensing products contaminated by thick clouds and shadows using multitemporal dictionary learning. IEEE Trans. Geosci. Remote Sens. **52**(11), 7086–7098 (2014)

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

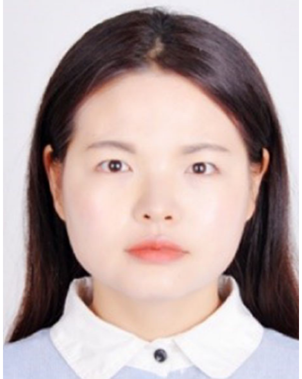


**Xinghe Yan** was born in Jiangsu, China, in 1993. He received the B.S. degree in automation from the School of Control Science and Engineering, Shandong University, Jinan, China, in 2016. He is currently working toward the M.S. degree in control science and engineering at the School of Control Science and Engineering, Shandong University, Jinan, China. His research interests include machine learning, deep learning and salient object detection.

**Zhenxue Chen** was born in Shandong, China, in 1977. He received the B.S. degree in automatic from School of Electrical Engineering and Automation at Shandong Institute of Light Industry, Jinan, China, in 2000, the M.S. degree in computer science from School of Information Science and Engineering at Wuhan University of Science and Technology, Wuhan, China, in 2003, and the Ph.D. degree in pattern recognition and intelligent systems from Institute of Image Recognition and Artificial Intelligence at Huazhong University of Science and Technology, Wuhan, China, in 2007. From 2012 to 2013, he was a visiting scholar with the Michigan State University, East Lansing, Michigan, USA. He is currently a professor with the School of Control Science and Engineering, Shandong University. His main areas of interest include image processing, pattern recognition and computer vision, with applications to face recognition. He has published over 100 papers in refereed international leading journals/conferences such as IEEE T-II, IEEE T-CSVT, IEEE T-IFS, IEEE T-VT, IEEE T-ITS, Information Sciences, Neurocomputing, Neural Computing and Applications, and SP-IC.

**Q. M. Jonathan Wu** (M'92-SM'09) received the Ph.D. degree in electrical engineering from the University of Wales, Swansea, UK, in 1990. He was with the National Research Council of Canada for ten years from 1995, where he became a senior research officer and a group leader. He is currently a professor with the Department of Electrical and Computer Engineering, University of Windsor, Windsor, ON, Canada. He has published more than 300 peer-reviewed papers in computer vision, image processing, intelligent systems, robotics and integrated microsystems. His current research interests include machine learning, 3-D computer vision, video content analysis, interactive multimedia, sensor analysis and fusion and visual sensor networks. He holds the Tier 1 Canada Research Chair in Automotive Sensors and Information Systems. He was the associate editor for IEEE Transactions on Systems, Man, and Cybernetics Part A, and the International Journal of Robotics and Automation. Currently, he is an Associate Editor for the IEEE Transactions on Neural Networks and Learning Systems and the journal of Cognitive Computation. He has served on technical program committees and international advisory committees for many prestigious conferences.

**Mengxu Lu** was born in Jiangsu, China, in 1997. She received the B.S. degree in automation from the School of Control Science and Engineering, Shandong University, Jinan, China, in 2019. She is currently working toward the M.S. degree in control science and engineering at the School of Control Science and Engineering, Shandong University, Jinan, China. Her research interests include machine learning, deep learning and semantic segmentation.



**Luna Sun** was born in Henan, China, in 1996. She received the B.S. degree in School of Automation from Jiangnan University, Wuxi, China, in 2019. She is pursuing the M.S. degree in control science and engineering at the School of Control Science and Engineering at the School of Control Science and Engineering, Shandong University, Jinan, China. Her current research interests include machine learning, deep learning and salient object detection.