IEICE TRANS. INF. & SYST., VOL.E103–D, NO.4 APRIL 2020

LETTER

Salient Region Detection with Multi-Feature Fusion and Edge Constraint

Cheng XU[†a)], ***Member***, **Wei HAN**[†], **Dongzhen WANG**[†], ***and*** **Daqing HUANG**[†], ***Nonmembers***

SUMMARY In this paper, we propose a salient region detection method with multi-feature fusion and edge constraint. First, an image feature extraction and fusion network based on dense connection structure and multi-channel convolution channel is designed. Then, a multi-scale atrous convolution block is applied to enlarge reception field. Finally, to increase accuracy, a combined loss function including classified loss and edge loss is built for multi-task training. Experimental results verify the effectiveness of the proposed method.

key words: *salient region detection, convolutional neural network, multi-feature fusion, edge constraint*

1. Introduction

Salient region detection is a basic research field in image processing and computer vision. Its purpose is to find the most salient region in an image. Salient region detection is widely used in various image processing and recognition tasks, including target location, visual tracking and semantic segmentation.

Traditional salient region detection usually uses the middle- and low-level features of images, such as regional texture, color, and contrast. Through feature extraction, the distinction between saliency and non-saliency is achieved [1]–[3]. These methods lack sufficient semantic information, so they do not work well in complex scenarios.

In recent years, deep learning has greatly promoted the development of salient region detection technology. Compared with the traditional algorithm, the accuracy of the detection results of the saliency model based on deep learning on the public dataset far exceeds those of the traditional algorithm. Wang et al. [4] present a salient region detection algorithm by integrating both local estimation and global search. Li et al. [5] proposed a framework by integrating the CNN-based saliency model with a spatial coherence model and multi-level image segmentations. Zhao et al. [6] designed a multi-context deep learning framework for salient object detection, including both global context and local context. Compared with traditional salient region detection methods, these methods greatly improve the detection accuracy, but because the pooling operation of a deep network cannot better preserve the edge information of the salient object, the edge of the salient object is usually blurred and the shape of the salient object is difficult to determine.

To solve this problem, a salient region detection method combining multi-feature fusion and edge constraint is designed in this paper, which focusing on how to effectively fuse multi-feature maps and overcome edge blurring in salient regions through supervised learning of edge information. Without any post-processing, this method can generate an accurate saliency map with precise boundaries.

In summary, the main contributions of this work are the following:

(1) The dense connection structure and multi-convolution channel are used to construct a multi-scale feature-fusion network with feature reuse.

(2) A multi-scale atrous convolution block (MACB) is constructed to extract features in larger receptive fields.

(3) The edge constraint on the salient object is added, with the classification of the saliency/non-saliency regions, the combined loss function is built, and the multi-task training method is used to enhance the edge-distinction ability of networks.

2. Methods

2.1 Network Framework

The network structure is shown in Fig. 1. The network consists of two sub-networks: a feature fusion network and salient region detection network. The backbone network of the feature fusion network uses DenseNet [9]. DenseNet enables each layer to connect directly with all its subsequent layers and features can be reused, so as to reduce the number of parameters and keep the network efficient. On this basis, we used the idea of feature pyramid networks (FPNs) [10] to output feature maps from Dense Block 2, Dense Block 3, and Dense Block 4 separately, and designed a multi-scale feature fusion block (MFFB). This block can fuse the shallow features with the adjacent deep features by adding the parallel convolution channel, in order to improve the semantic strength of the shallow features. After the feature fusion network executes, feature maps of three different scales can be obtained and sent to the salient region detection network for further feature extraction and pixel classification. The salient region detection network includes a feature extraction block (FEB) and an MACB. The skip-connection structure of ResNet is adopted in the FEB, the main function of which is to further extract features and adjust the scale of the feature maps. The MACB employs atrous con-

Manuscript received October 6, 2019.
Manuscript revised December 27, 2019.
Manuscript publicized January 17, 2020.

[†]The authors are with the Nanjing University of Aeronautics and Astronautics, China.
a) E-mail: xc88@vip.qq.com (Corresponding author)

DOI: 10.1587/transinf.2019EDL8181

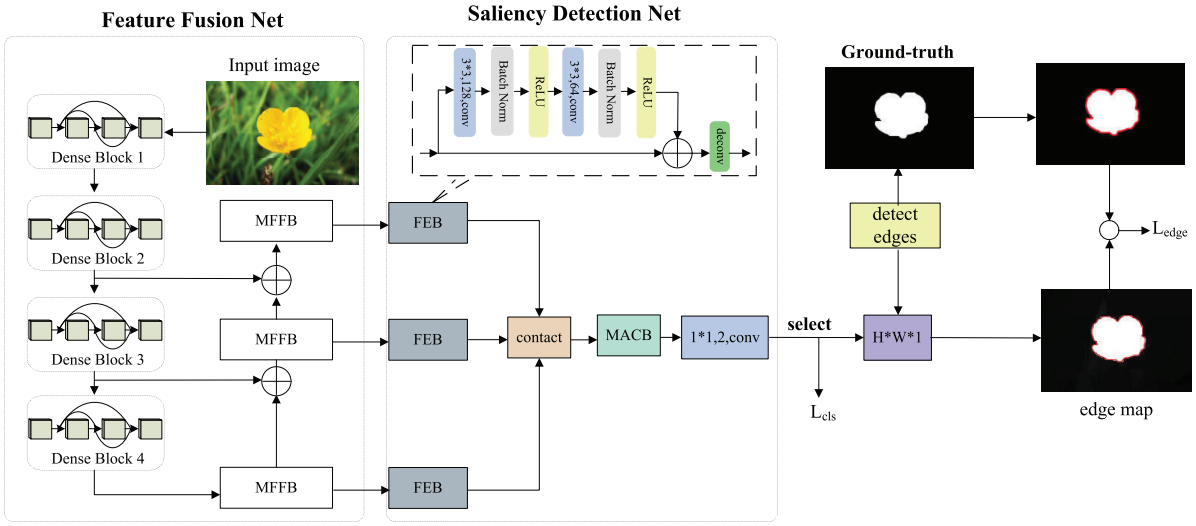

Fig. 1 Architecture overview of the proposed network.

volution with a rate = 6, 12, and 18 to extract salient features from the multi-scale feature maps and uses 1*1 convolution to reduce the dimension of the feature maps. Then, the cross-entropy function is used as classified loss to classify the saliency/non-saliency pixels. After output classification, the salient class is selected, and then edge detection is performed on the salient map and ground-truth map separately, and the edge loss representing the edge constraint can be calculated. We combine classified loss and edge loss in a multi-task learning network, and use the correlation between tasks to promote learning. Finally, an accurate saliency map with precise boundaries is generated.

2.2 Multi-Scale Feature Fusion Block

Shallow features usually do not have sufficient ability to express features, while deep features lack good ability to describe details. In this case, we use residual structure and deconvolution to form a multi-scale feature fusion block. The structure of this block is shown in Fig. 2. The MFFB is divided into three branches: branch 1, branch 2, and branch 3. Branch 1 adopts a 1*1 convolution kernel as the skip connection, which can alleviate the problem of gradient disappearance. To enhance the expressive ability of shallow features, we use a 3*3 convolution core as branch 2. Branch 2 has a larger field of perception, which is convenient for capturing image details; branch 3 consists of 3*3 convolution and deconvolution. Branch 3 uses the deeper features close to it, enhances the expression ability of the shadow features, and makes it easy to distinguish the salient object from the background. When MFFB is used to extract the features of Dense Block 4, as shown in Fig. 1, branch 3 is removed because no deeper features can be used for fusion.

2.3 Feature Extraction Block

The structure of FEB is shown in Fig. 1. The skip connection is adopted in this structure, and two 3*3 convolution including batch-norm layers and ReLu layers are used in the branch to extract detail features, which are fused with the original channel features to achieve the effect of distinguishing fine features. Finally, the feature map is adjusted to the same size by deconvolution.

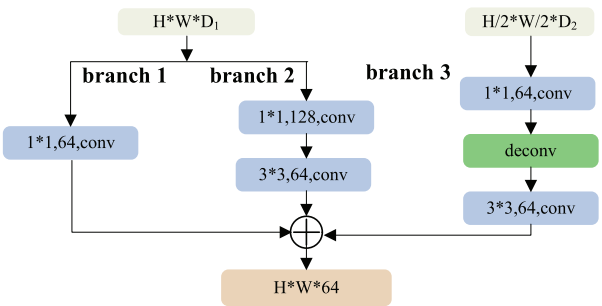

Fig. 2 Multi-scale feature fusion block.

2.4 Multi-Scale Atrous Convolution Block

Traditional convolution limits the size of receptive field, so it is necessary to use the pooling layer to enlarge the receptive field by dimension reduction of the feature map. The pooling dimension reduction will miss the details of the image, resulting in irreversible loss. Therefore, to extract more salient features, a MACB is designed by using atrous convolution [8]. An image I is input and two-dimensional convolution kernels C is used to perform a pixel-by-pixel convolution operation. For the general convolution operation, we have

$$f(i,j) = \sum_m \sum_n I(i+m, j+n) C(m,n) \tag{1}$$

The rate r is added in the atrous convolution. Then,

$$f(i,j) = \sum_m \sum_n I(i+m*r, j+n*r) C(m,n) \tag{2}$$

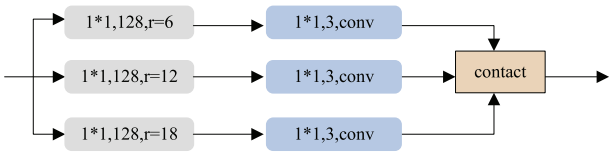

Fig. 3 Multi-scale atrous convolution block.

when $r \geq 2$, the receptive field can be enlarged without changing the size of the convolution core and adding additional parameters. The receptive field can be expressed as

$$R_{field} = [(k+1) * (r-1) + k]^2 \tag{3}$$

where k represents the size of the convolution core. To obtain features of different scales, the feathers are extracted by atrous convolution with $r = 4$, 8, and 16, as shown in Fig. 3.

2.5 Loss

In the proposed method, an edge constraint is added to enhance the clarity of the edge of salient objects. Therefore, the loss function can be divided into classified loss and edge loss.

(1) Classified loss

The parameters of the network can be optimized by minimizing the errors between the predicted values and corresponding annotations. We use a binary classification network to segment salient objects and non-salient background regions. The goal of the network optimization is achieved by minimizing the cross-entropy of the softmax classifier:

$$L_i^{cls} = -\frac{1}{N}\sum_{k=1}^{N}(y_k^{cls}\log(p_k) + (1-y_k^{cls})(1-\log(p_k))) \tag{4}$$

where p_k is the probability produced by the network that indicates pixel x_k in the salient region. The notation $y_k^{cls} \in \{0, 1\}$ denotes the ground-truth label. N is the number of pixels in the image.

(2) Edge loss

Edge loss uses the predicted and the matched ground-truth masks as input, which are edge detected with edge detection filter Sobel. Afterwards, the difference between the predicted and ground-truth edge maps are determined. For this task, we choose Euclidean loss, which can express the absolute difference between the predicted and ground-truth edge maps.

$$L_i^{edge} = \frac{1}{N}\sum_{k=1}^{N}\|y_k - \hat{y}_k\|^2 \tag{5}$$

where $\hat{y}_k$ is the pixel of edge map predicted by the network, and y_k the pixel of the ground-truth edge map. N is the number of pixels in the image.

(3) Multi-task training

There are two loss functions in the whole network, including one classification loss and one edge loss. We adopted the strategy of joint training and designed a multi-task loss function, in which each output loss can be used as the regularization term of other losses to prevent overfitting to some extent. During training, the multi-task loss function can provide additional gradient signals to prevent gradient disappearance. The multi-task loss function can be expressed as follows:

$$\min \sum_{i=1}^{M}\sum_{j\in\{cls,edge\}}\omega_j L_i^j \tag{6}$$

where M is the number of training samples, ω_j the weight coefficient.

Table 1 Quantitative performance on three benchmark datasets.

Algorithms	ASD		PASCAL-S		ECSSD	
	F_β	MAE	F_β	MAE	F_β	MAE
GMR	0.911	0.077	0.662	0.219	0.741	0.192
DSR	0.887	0.082	0.646	0.207	0.734	0.175
HDCT	0.885	0.121	0.607	0.229	0.707	0.198
LEGS	0.904	0.062	0.751	0.155	0.831	0.121
MDF	0.931	0.052	0.758	0.144	0.830	0.106
MCDL	0.928	0.035	0.736	0.144	0.839	0.103
SSD	0.932	0.035	0.762	0.121	0.837	0.102
Ours	0.933	0.034	0.773	0.119	0.869	0.081

3. Experiment

We compared the proposed algorithm with six salient region detection algorithms, including GMR [1], DSR [2], HDCT [3], LEGS [4], MDF [5], MCDL [6] and SSD [7]. GMR, DSR and HDCT are traditional detection algorithms using low-level features, while LEGS, MDF, MCDL and SSD are salient region detection algorithms based on deep learning. We obtained the result images from the project site of each algorithm. The results which were not provided were generated from the authors' source codes published in the web. We used three typical test datasets: ASD, PASCAL-S, and ECSSD for our evaluation.

Precision-recall (PR) curves, F-measure (F_β) and mean absolute error (MAE) were used as metrics to evaluate the performance of salient region detection. The PR curve is computed by binarizing the saliency maps under different probability thresholds ranging from 0 to 1 and comparing against the ground truth. The comparisons on PR graph is presented in Fig. 4. Maximum F_β scores and MAE values are also described in Table 1.

As shown in Fig. 4 and Table 1, our model achieves superior quantitative max F_β, MAE and PR performance across the board when compared to GMR, DSR, HDCT, LEGS, MDF, MCDL and SSD. This is because our model fuses multiple feature maps through a feature fusion network to better describe the semantics and details of images, and then optimizes salient maps through an edge constraint, which improves the salient map's accuracy.

Figure 5 is visual comparison of salient region detection results with and without the edge loss term. When dealing with complex scenes, the proposed algorithm can

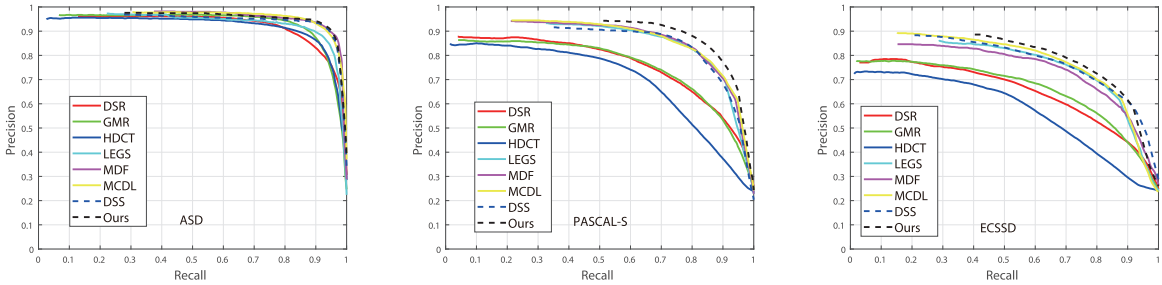

Fig. 4 Precision-recall curves on the three benchmark datasets.

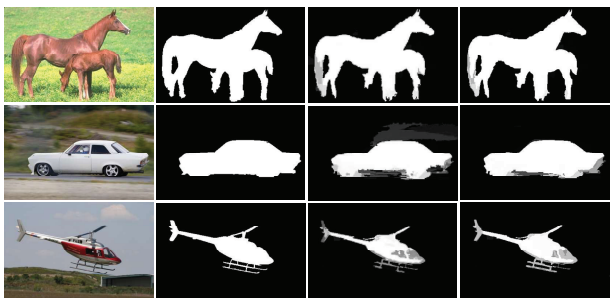

Fig. 5 Visual comparison of salient region detection results with and without the edge loss term.

not only smoothly and uniformly display salient regions, but also shows a good effect in edges. The salient maps obtained are clear as a whole, and the salient object and background are clearly distinguished.

In order to reflect the influence of MACB and edge constraint on the performance of the algorithm, we have carried out a comparative experiment on ECSSD dataset. After adding the MACB to the backbone network, F_β increased by 9.2% and MAE decreased by 12.7%. Further adding edge constraint, F_β increased by 7.5% and MAE decreased by 9.7%. It can be seen that MACB and edge constraint significantly improve the performance of salient region detection.

4. Conclusion

To solve the problem of blurred-edge representation of salient objects in complex backgrounds, we propose a salient region detection method in this letter. By extracting and fusing multi-scale features of images, image details and semantic information are effectively utilized. Combined with the edge loss, the edge information of salient objects can be captured by the network. The effectiveness of the proposed algorithm is compared with seven salient region detection algorithms that have been utilized in recent years. The results show that the proposed algorithm can detect salient objects more accurately, and can better represent the information of salient objects, and with clearer boundaries.

Acknowledgments

This work is supported by National Natural Science Foundation of China (no.61601222), Jiangsu Provincial Natural Science Foundation of China (no.BK20160789), China Postdoctoral Science Foundation (no.2018M632303).

References

[1] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," Proc. 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '13, pp.3166–3173, 2013.

[2] X. Li, H. Lu, L. Zhang, X. Ruan, and M.-H. Yang, "Saliency detection via dense and sparse reconstruction," Proc. 2013 IEEE International Conference on Computer Vision, ICCV '13, pp.2976–2983, 2013.

[3] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform," Proc. 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '14, pp.883–890, 2014.

[4] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '15, pp.3183–3192, 2015.

[5] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '15, pp.5455–5463, 2015.

[6] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," Proc. 2015 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '15, pp.1265–1274, 2015.

[7] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '17, pp.5300–5309, 2017.

[8] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A.L. Yuille, "Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs," IEEE Trans. Pattern Anal. Mach. Intell., vol.40, no.4, pp.834–848, 2018.

[9] G. Huang, Z. Liu, L.V.D. Maaten, and K.Q. Weinberger, "Densely connected convolutional networks," Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '17, pp.2261–2269, 2017.

[10] T.-Y. Lin, P. Dollár, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," Proc. 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR '17, pp.936–944, 2017.