Making BBS-Net More Efficiency for RGB-D Salient Object Detection

Qiaoling Liu, Xiaoying Guo, Yongquan Li, Yan Xu*

College of Electronic and Information Engineering, Shandong University of Science and Technology, Qingdao, Shandong, 266000, China

*Corresponding author’s e-mail: xuyan@sduet.edu.cn

Abstract. BBS-Net is a novel bifurcated backbone strategy network based on cascaded refinement, which achieves the most advanced performance in RGB-D salient object detection models. However, the performance of the model in multi-scale feature extraction and depth feature enhancement is not strong enough, and the generated saliency maps contain a lot of background distraction and blurred edges. In this paper, based on BBS-Net framework, we adopt more effective feature extraction network and depth feature enhancement module. Specifically, 1) Use the Res2Net-50 structure that constructs hierarchical residual-like connections within a single residual block as a feature extraction network; 2) Use ECA module as channel attention to extract depth information and enhance cross-modal compatibility. The comprehensive experiments of our method on 6 datasets show that it outperforms the original BBS-Net and 7 current state-of-the-art methods in 4 evaluation metrics.

1. Introduction

In recent years, salient object detection (SOD) based on RGB-D has received widespread attention and various models have been proposed. BBS-Net[1] is the state-of-the-art performance model in RGB-D SOD models. It is a bifurcated backbone strategy network using a cascaded refinement strategy. The features extracted by ResNet-50[2] are divided into two parts: high-level features and low-level features. First, a cascaded decoder is used to aggregate the high-level features to generate the initial saliency map. Low-level features are refined by an element-wise multiplication with the initial saliency map. Finally, the refined low-level features integrated by another cascaded decoder to predict the final saliency map.

Although BBS-Net has achieved encouraging results, the performance of the model is not strong enough. It can be seen from the saliency maps in the fourth column of Fig. 1 that there are still many background distraction and edge blur problems in the maps, and complete salient objects cannot be detected. We believe that there are two major challenges:
Fig. 1 Saliency maps of BBS-Net

(1) Suboptimal feature extraction network.

Objects may appear with different size in depth maps and RGB images, and perceiving information of different scales is important for understanding objects. The network can detect objects more accurately through context information. As a classic feature extraction network, ResNet-50 uses repeated 1×1, 3×3, and 1×1 convolutions to extract features. Its performance in multi-scale feature extraction is not strong enough, which makes the feature expression insufficient.

(2) The original depth-enhanced module has poor effect.

Depth maps are much noisier and textureless than RGB images, which may introduce feature noise and redundancy into the network. Through deep analysis of the saliency maps generated by the original model and the performance of each module, we found that the performance of the depth enhancement module is not effective enough, which will affect the compatibility of multi-modal features and the extraction of depth informative cues.

In computer vision tasks, obtaining multi-scale representations require the convolutional filters to use a large range of receptive fields to describe objects of different scales. Convolutional neural networks (CNNs) gradually learn multi-scale features from coarse to fine through a stack of convolutional filters. Here, we specifically use Res2Net-50[3] as the backbone network to extract features. It expresses multi-scale features at a granular level by constructing hierarchical residual-like connections within the residual block.

Compared with the original network, we specifically use ECA-Net (i.e., ECA module)[4] to realize the connection between depth channel features. ECA-Net is a channel attention that uses a local cross-channel interaction strategy. The model uses 1D convolution to be more efficient, and the coverage of interaction can be adaptively selected according to the number of channels.

We have verified the effectiveness of our model via extensive studies and comparisons. In summary, our main contributions are as follows:

(1) In order to more fully extract the depth and RGB features, we use the Res2Net-50 as the backbone network for multi-scale feature extraction. This feature extraction network can extract features at a granular level and increase the receptive field, enhancing the representation of features.

(2) To more effectively enhance the depth features and improve the compatibility of RGB and depth features, we use ECA-Net as a more effective channel attention module. This channel attention uses 1D convolution to achieve local cross-channel interaction and enhance the connection between depth channels.

(3) Without bells and whistles, our model outperforms the original BBS-Net network and 7 state-of-the-art alternatives, over 6 widely used benchmark datasets.

2. Methods

For the specific improvement methods of BBS-Net, we first introduce the purpose of adopting the new feature extraction network and the basic principle of Res2Net module in 2.1, and then introduce the reason and structure of ECA module as our channel attention in 2.2.
2.1. Feature extraction network

By replacing the feature extraction network in the original BBS-Net network with the Res2Net-50, the purpose is to enhance the model's multi-scale feature extraction capabilities, expand the receptive field of the feature map, and capture richer context information to further improve the feature representation ability of the model.

Res2Net module is a further improvement made on the basis of ResNet block, which constructs hierarchical residual-like connections within a single residual block. The structure diagram of Res2Net module is shown as in Fig. 2.

Within a single residual block, the feature map after 1×1 convolution is evenly divided into groups according to the number of channels, and each group uses 3×3 convolution for feature extraction. The features after 3×3 convolution are added with the original features of the next group. In order to reduce the parameters of the module, the first group omits the 3×3 convolution, which can also be regarded as the reuse of features. Finally, the feature maps obtained in each group are spliced, and information fusion of different scales is carried out through a 1×1 convolution. The expression of the Res2Net module can be written as:

\[
y_i = \begin{cases} 
x_i & i = 1; 
\text{Conv}(x_i) & i = 2; 
\text{Conv}(x_i + y_{i-1}) & 2 < i \leq s.
\end{cases}
\]  

Where \(x_i\) indicates a subset of the feature map, \(\text{Conv}\) is a 3×3 convolution operation, \(y_i\) and \(y_{i-1}\) indicate the final output of the corresponding feature subsets. It can be seen from the expression that every 3×3 convolution will increase the receptive field of the feature map, and finally a multi-scale feature map is spliced.

2.2. More effective channel attention

Depth maps captured from state-of-the-art sensors are usually poor, it is very important to use a high-performance attention module to extract useful information from the depth features and improve multi-modal compatibility. Using ECA module as our new channel attention can enable the model to greatly reduce the amount of parameters while strengthening the connection between deep feature channels, and enhance the detection performance of the model.

ECA module adopts a local cross-channel strategy on the basis of SE block[5]. The input feature maps first aggregate global features through global average pooling, then adjust the channel weight through convolution operations, and finally multiply the weight value with the original feature maps. The model structure is shown in Fig.3.
The model only considers the information of the current channel and its $k$ adjacent channels, so it has only $k \times C$ parameters. In addition, channels sharing the same learning parameters can be used to further improve model efficiency. The expression of the $i$th value of feature $y$ is:

$$\omega_i = \text{Sigmoid} \left( \sum_{j=1}^{k} \alpha^j_i y^j_i \right), y^j_i \in \Omega^k_i$$

In which $\omega_i$ indicates the output of $y_i$ after weight operation, $\alpha^j_i$ is the weight of $y_i$ or one of its neighbors, and $\Omega^k_i$ represents the set of $y_i$ and its $k$ neighbors.

In addition, ECA module can adaptively select the coverage of interaction according to the number of channels. The calculation formula is:

$$k = \Psi(C) = \left\lfloor \frac{\log_2(C)}{y} + b \right\rfloor$$

Where $k$ indicates the size of the adaptive convolution kernel, $C$ is the number of channels. $y$ and $b$ are parameters, and it is recommended to set to 2 and 1 respectively.

3. Experiments

3.1. Experimental Settings

Implementation Details. Our model uses the PyTorch framework to experiment on a single NVIDIA 3080 GPU. Parameters of the Res2Net-50 are initialized from the model pretrained on ImageNet, and other parameters are initialized as the default settings of PyTorch. We train the model with a batch size of 8 for 200 epochs, using the Adam optimization algorithm to optimize our model, the initial learning rate is set to le-4, and it drops 10 times every 60 epochs. The methods used for data enhancement include random flip, border clip and random rotation.

Datasets. We used 1485 images from the NJU2000[6] and 700 images from the NLPR[7] for training. The remaining samples of NJU2000 and NLPR and the entire STERE[8], DES[9], LFSD[10], SSD[12], and SIP[11] datasets are used for testing.

Evaluation Metrics. We adopt 4 commonly used metrics, namely S-measure, maximum F-measure (maxF), maximum E-measure (maxE) and absolute error (MAE).

3.2. Ablation Experiments and Analyses

In this section, we explore the impact of different modules in the improved method on the datasets.

New feature extraction network. To prove the effectiveness of Res2Net-50 as the backbone network, we compare the results of Res2Net-50 as the backbone with the results of the original model. As shown in Tab.2, we could see that the use of the new feature extraction network brings significant improvements to the results.
Tab.1 Effectiveness analysis of Res2Net-50. ↑ (↓) denotes that the higher (lower) the better

| Model                  | NJU2000[6] | NLPR[7] | STERE[8] | DES[9] | SIP[11] |
|------------------------|------------|---------|----------|--------|---------|
|                        | S↑ | MAE↓ | S↑ | MAE↓ | S↑ | MAE↓ | S↑ | MAE↓ | S↑ | MAE↓ |
| BBS-Net                | 0.921 | 0.035 | 0.930 | 0.023 | 0.908 | 0.041 | 0.933 | 0.021 | 0.879 | 0.055 |
| BBS-Net-Res2Net        | 0.927 | 0.031 | 0.931 | 0.023 | 0.911 | 0.040 | 0.933 | 0.020 | 0.898 | 0.045 |

More effective channel attention. We replaced the original channel attention with ECA module and retrained. The experimental results are shown in Tab.3. Numerically, ECA module is more effective than the original channel attention, indicating that the new channel attention enhances the extraction of depth features and helps cross-modal feature compatibility.

Tab.2 Effectiveness analysis of ECA module

| Model                  | NJU2000[6] | NLPR[7] | LFSD[10] | DES[9] | SIP[11] |
|------------------------|------------|---------|----------|--------|---------|
|                        | S↑ | MAE↓ | S↑ | MAE↓ | S↑ | MAE↓ | S↑ | MAE↓ | S↑ | MAE↓ |
| BBS-Net                | 0.921 | 0.035 | 0.930 | 0.023 | 0.864 | 0.072 | 0.933 | 0.021 | 0.879 | 0.055 |
| BBS-Net-ECA            | 0.924 | 0.034 | 0.931 | 0.023 | 0.868 | 0.069 | 0.934 | 0.021 | 0.887 | 0.051 |

3.3. Compare with the State-of-the-art

We compare our model with BBS-Net and 7 deep learning based salient object detection models including DF[12], AFNet[13], MMCI[14], PCF[15], TANet[16], CPFP[17], D3Net[11].

Quantitative Results

As shown in Tab.3, our model outperforms the state-of-the-art methods on most evaluation metrics contain S-measure, maxF, maxE and MAE.

Tab.3 Quantitative comparison of models including S-measure, maxF, maxE, and MAE on six datasets

| Dataset | Metric | DF[12] | AFNet[13] | MMCI[14] | PCF[15] | TANet[16] | CPFP[17] | D3Net[11] | BBS-Net | Ours |
|---------|--------|--------|-----------|----------|---------|-----------|----------|-----------|---------|------|
| NJU2000[6] | S-measure↑ | 0.763 | 0.772 | 0.858 | 0.877 | 0.878 | 0.879 | 0.900 | 0.921 | 0.929 |
|          | maxF↑ | 0.804 | 0.775 | 0.852 | 0.872 | 0.874 | 0.877 | 0.900 | 0.920 | 0.931 |
|          | maxE↑ | 0.864 | 0.853 | 0.915 | 0.924 | 0.925 | 0.926 | 0.926 | 0.939 | 0.949 |
|          | MAE↓ | 0.141 | 0.100 | 0.079 | 0.059 | 0.060 | 0.053 | 0.046 | 0.035 | 0.030 |
| NLPR[7]  | S-measure↑ | 0.778 | 0.771 | 0.815 | 0.841 | 0.863 | 0.867 | 0.897 | 0.918 | 0.921 |
|          | maxF↑ | 0.880 | 0.879 | 0.913 | 0.925 | 0.941 | 0.932 | 0.953 | 0.961 | 0.962 |
|          | maxE↑ | 0.085 | 0.058 | 0.059 | 0.044 | 0.041 | 0.036 | 0.030 | 0.023 | 0.023 |
|          | MAE↓ | 0.141 | 0.075 | 0.068 | 0.064 | 0.060 | 0.051 | 0.066 | 0.041 | 0.038 |
| STERE[8] | S-measure↑ | 0.757 | 0.825 | 0.873 | 0.875 | 0.871 | 0.879 | 0.835 | 0.908 | 0.915 |
|          | maxF↑ | 0.847 | 0.887 | 0.927 | 0.925 | 0.923 | 0.925 | 0.911 | 0.924 | 0.947 |
|          | maxE↑ | 0.141 | 0.075 | 0.068 | 0.064 | 0.060 | 0.051 | 0.066 | 0.041 | 0.038 |
|          | MAE↓ | 0.093 | 0.068 | 0.065 | 0.049 | 0.046 | 0.038 | 0.030 | 0.021 | 0.019 |
| DES[9]   | S-measure↑ | 0.752 | 0.770 | 0.848 | 0.842 | 0.858 | 0.872 | 0.900 | 0.933 | 0.940 |
|          | maxF↑ | 0.766 | 0.728 | 0.822 | 0.804 | 0.827 | 0.846 | 0.888 | 0.927 | 0.936 |
|          | maxE↑ | 0.870 | 0.881 | 0.929 | 0.939 | 0.910 | 0.923 | 0.943 | 0.966 | 0.973 |
|          | MAE↓ | 0.093 | 0.068 | 0.065 | 0.049 | 0.046 | 0.038 | 0.030 | 0.021 | 0.019 |
| LFSD[10] | S-measure↑ | 0.783 | 0.738 | 0.787 | 0.786 | 0.801 | 0.828 | 0.825 | 0.864 | 0.875 |
|          | maxF↑ | 0.817 | 0.744 | 0.771 | 0.775 | 0.796 | 0.826 | 0.810 | 0.859 | 0.866 |
|          | maxE↑ | 0.857 | 0.815 | 0.839 | 0.827 | 0.847 | 0.863 | 0.862 | 0.901 | 0.912 |
|          | MAE↓ | 0.145 | 0.133 | 0.132 | 0.119 | 0.111 | 0.098 | 0.095 | 0.072 | 0.064 |
| SIP[11]  | S-measure↑ | 0.653 | 0.729 | 0.833 | 0.842 | 0.835 | 0.850 | 0.860 | 0.879 | 0.889 |
|          | maxF↑ | 0.657 | 0.712 | 0.818 | 0.838 | 0.830 | 0.851 | 0.861 | 0.883 | 0.899 |
|          | maxE↑ | 0.759 | 0.819 | 0.897 | 0.901 | 0.895 | 0.903 | 0.909 | 0.922 | 0.928 |
|          | MAE↓ | 0.185 | 0.118 | 0.086 | 0.071 | 0.073 | 0.064 | 0.063 | 0.055 | 0.048 |
Visual Comparison

In Fig.4, we show some visualization results. These images are divided into five challenging situations, including simple scene, complex scenes, small object, multiple objects, and low contrast. As shown in Fig.4, we give a simple example in the first row. All the methods are basically good in performance. Two small object examples are displayed in rows 2-3. It is obvious that other methods have different degrees of background distraction when detecting people. The multiple objects situations are shown in rows 4-5. Our method can locate all salient objects in both examples and the segmentation effects are the best. Although the quality of the depth map in line 5 is relatively poor, our model can still detect all objects. In 6th-7th rows, we show the complex scene. Compared with other methods, our model does not mistakenly treat background distraction as salient objects. Besides, we also sample some images (8th-9th rows) whose scene is low contrast. Although the contrast of RGB images is very low, our method can effectively use the information in the depth map to detect reliable results.

Acknowledgments
This work was supported by the Natural Science Foundation of China (11547037, 11604181), Shandong Graduate Education Innovation Project (01040105305), Hisense Group research and development Center Project, and the Scholarship Fund of SDUST.

References
[1] Fan, D.P., Zhai, Y.J., Borji, A., Yang, J., Shao, L. (2020) BBS-Net: RGB-D Salient Object Detection with a Bifurcated Backbone Strategy Network. In: European Conference on Computer Vision. Glasgow. pp. 275-292.
[2] He, K., Zhang, X., Ren, X., and Sun, J. (2016) Deep Residual Learning for Image Recognition. In: Computer Vision and Pattern Recognition. Las Vegas. pp. 770–778.
[3] Gao, S.H., Cheng, M.M., Zhao, K., Zhang, X.Y., Yang, M.H., Torr, P. (2019) Res2net: A new multi-scale backbone architecture. IEEE Transactions on Pattern Analysis and Machine Intelligence, 43(2): 652-662.
[4] Wang, Q., Wu, B., Zhu, P., Li, P., Zuo, W. and Hu, Q. (2020) ECA-Net: Efficient Channel Attention for Deep Convolutional Neural Networks. In: Computer Vision and Pattern Recognition. Seattle. pp. 11531-11539.
[5] Hu, J., Shen, L., Albanie, S., Sun, G. and Wu, E. (2018) Squeeze-and-Excitation Networks. In:
[6] Ju, R., Ge, L., Geng, W., Ren, T., Wu, G. (2014) Depth saliency based on anisotropic center-surround difference. In: International Conference on Image Processing. Paris. pp. 1115–1119.

[7] Peng, H., Li, B., Xiong, W., Hu, W., Ji, R. (2014) RGBD Salient Object Detection: A Benchmark and Algorithms. In: European Conference on Computer Vision. Zurich. pp. 92–109.

[8] Niu, Y., Geng, Y., Li, X., Liu, F. (2012) Depth Enhanced Saliency Detection Method. In: Computer Vision and Pattern Recognition. Providence. pp. 454–461.

[9] Cheng, Y., Fu, H., Wei, X., Xiao, J., Cao, X. (2014) Depth Enhanced Saliency Detection Method. In: International Conference on Internet Multimedia Computing and Service. Xiamen. pp. 23–27.

[10] Li, N., Ye, J., Ji, Y., Ling, H., Yu, J. (2014) Saliency Detection on Light Field. In: Computer Vision and Pattern Recognition. Columbus. pp. 2806-2813.

[11] Fan, D. P., Lin, Z., Zhang, Z., Zhu, M., Cheng, M. M. (2020) Rethinking RGB-D Salient Object Detection: Models, Data Sets, and Large-Scale Benchmarks. IEEE Transactions on Neural Networks and Learning Systems.

[12] Qu, L., He, S., Zhang, J., Tian, J., Tang, Y., Yang, Q. (2017) RGBD Salient Object Detection via Deep Fusion. IEEE Transactions on Image Processing, 26(5):2274-2285.

[13] Wang, N., Gong, X. (2019) Adaptive Fusion for RGB-D Salient Object Detection. IEEE Access, 7: 55277-55284.

[14] Chen, H., Li, Y., Su, D. (2019) Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. Pattern Recognition, 86:376–385.

[15] Chen, H., Li, Y. (2018) Progressively Complementarity-Aware Fusion Network for RGB-D Salient Object Detection. In: Computer Vision and Pattern Recognition. Salt Lake City. pp. 3051–3060.

[16] Chen, H., Li, Y. (2019) Three-stream attention-aware network for RGB-D salient object detection. Transactions on Image Processing, 28(6): 2825–2835.

[17] Zhao, J.X., Cao, Y., Fan, D.P., Cheng, M.M., Li, X.Y., Zhang, L. (2019) Contrast Prior and Fluid Pyramid Integration for RGD Salient Object Detection. In: Computer Vision and Pattern Recognition. Long Beach. pp.3927–3936.