RGBD salient object detection based on depth feature enhancement

Yue Shi*, Wanjun Yu and Qianxun Guan
School of Computer Science and Information Engineering, Shanghai Institute of Technology, Shanghai, 201418, China
*Corresponding author’s e-mail: 206141144@mail.sit.edu.cn

Abstract: The depth map contains spatial location information, which has been proven to be beneficial for RGBD salient object detection. However, the quality of the depth map will directly affect the accuracy of RGBD saliency detection. In order to solve the problem of different quality of depth maps, a detection model based on depth enhancement is designed to enhance the available depth information and filter low-quality depth maps. By comparing four evaluation indicators with five advanced models, the experimental results show that the model in this paper is advanced.

1. Introduction
The salient object detection is to allow the computer to simulate human vision, and to select the most interesting area or object when observing a certain area. At present, salient object detection has been applied to many directions of computer vision, such as image segmentation, object detection, image captioning, and so on. In the past, many SOD models mostly used RGB images for saliency detection, but when it comes to complex scenes, the model often does not have a good detection effect. In recent years, with the rapid development of hardware technology and depth information acquisition equipment, researchers have begun to combine RGB images with depth maps for salient object detection. The depth map has rich spatial structure information, because each pixel value of it is the actual distance between the sensor and the object. This method of combining depth maps for detection is called RGBD saliency detection.

In order to effectively fuse RGB features and depth features, Han et al. [1] proposed a dual-stream model, which extracts RGB features and depth features separately, and then performs feature fusion in layers. Zhao et al. [2] designed a pyramid integration module to integrate the enhanced depth features with RGB features.

At present, although many RGBD saliency detection models have achieved good results, there are still many shortcomings. For example, most models do not process the depth information, directly fusing it with RGB features, without considering that the quality of the depth map is an important factor that affects the detection accuracy.

In response to the above problems, this paper designs a multi-level depth-enhanced RGBD saliency detection model. For high-quality depth maps, increase the feature weights. Conversely, for low-quality depth maps, reduce their weight to effectively filter depth maps with misleading information. Compared with the five recently released models, our model has higher detection performance on five widely used data sets.
2. The proposed method
The model structure of this article is as shown in the figure below. Before the feature fusion is performed, the depth map information is first extracted through the enhancement module and the invalid information is filtered. At the same time, in order to ensure that the model can effectively link the high and low layer information, the fourth and fifth layer side outputs are added at the pixel level, and the first, second, and third layer side outputs are added at the pixel level.

![Network structure](image1)

2.1 Network structure
The whole model refers to the idea of mid-term fusion, and sends RGB and Depth into two identical backbone networks. The backbone network used in this article is VGG16. According to the structure of the VGG16 network, the model is divided into five layers. Considering that the depth information extracted only through the backbone network has high redundancy information, there is an enhancement module immediately after each layer of the depth stream, and the depth information processed by the enhancement module is combined with the information of the RGB stream. Fusion processing. In order to ensure the reliability of local information and global information, the side outputs of different layers are correlated.

2.2 Implementation details

![Depth enhancement module](image2)

H*W*C is the image size and C is the number of channels, Rate is the expansion rate
Considering that the features of the depth map extracted by the backbone network are no longer limited to a single channel, and the features contained in the depth map are more global information, the model in this paper uses different expansion convolution rates in the enhancement module to extract the backbone network. The depth information is further processed.

As shown in the figure, the depth features extracted by the backbone network are first dilated and convolved twice, and then the channel and attention operations are performed on each convolution result. In order to ensure the effectiveness of the global information, the two dilated convolutions the result is pixel-added.

\[ f_{\text{depth}}^{l_1} = SA(CA(R_1)) \]  
\[ f_{\text{depth}}^{l_2} = SA(CA(R_2)) \]  

Among them, SA stands for spatial attention operation, CA stand for channel attention operation, and R1 and R2 respectively represent two different dilated convolution results.

\[ f_{\text{depth}}^l = \text{Conv}(f_{\text{depth}}^{l_1} \oplus f_{\text{depth}}^{l_2}, \omega) \]  

Conv stands for convolution operation, \( \oplus \) represents pixel addition operation, \( \omega \) is the convolution kernel parameter, here it represents the 3*3 convolution kernel, \( f_{\text{depth}}^l \) is the enhanced feature.

2.3 Training

Similar to the previous work [3, 4, 5], we use standard cross-entropy loss to supervise \( S_r \) and \( S_o \), the formula is:

\[ \text{LOSS} = \sum_{m=(r,c)} \lambda_m L_{\text{BCE}}(S_m, G) \]  

Among them, \( \text{LOSS} \) means total loss, \( L_{\text{BCE}} \) represents cross-entropy loss, \( G \) denotes the ground truth, and \( \lambda_m \) emphasizes each supervision. Experimental setting \( \lambda_c=1, \lambda_r=\lambda_d=0.5 \). Finally, \( S_c \) is the final forecast results.

3. The experimental results

We tested this model on five widely used RGBD data sets [6, 7, 8, 9], namely NLPR, SIP, DUT-RGBD, STERE and RGBD135. In order to prove the effectiveness of this method, compare it with Five advanced models were compared, including CPFP[2], CMW[9], D3NET[7], ICNET[10], DCMF[11]. Four evaluation indicators were used for evaluation, namely maximum F-measure (Fmaxβ), S-measure \( S(\alpha) \), maximum E-measure (Emax ξ) and mean absolute error (MAE).

1) The weighted F-measure offers a unified solution to the evaluation of non-binary and binary maps. And its formula is defined as:

\[ F_{\beta}^{\omega} = \frac{(1+\beta^2) \times \text{Precision}^\omega \times \text{Recall}^\omega}{\beta^2 \times \text{Precision}^\omega + \text{Recall}^\omega} \]  

Where \( \beta^2 \) is set to 0.3.

2) S-measure is the recently proposed structural similarity measure in the binary map evaluation field. And its formula is defined as:

\[ S = \alpha \times S_o + (1 - \alpha) \times S_r \]  

\( S_r \) represents region perception, \( S_o \) represents object perception, where \( \alpha \) is the balance parameter and set to 0.5 in this paper.

3.1 Table and visual results

This paper makes a quantitative comparison between the proposed method and the latest algorithm through tables and visual comparison through images, as shown below.
Table 1. Index data of each model result

| Models | SIP | NLPR |
|--------|-----|------|
|        |     | F    | avgF | MAE | S   | F   | avgF | MAE | S   |
| FCPFP  | 0.8189 | 0.8209 | 0.0636 | 0.8501 | 0.8227 | 0.8405 | 0.0358 | 0.8884 |
| CMW    | 0.8511 | 0.8510 | 0.0620 | 0.8674 | 0.8593 | 0.8774 | 0.0293 | 0.9171 |
| D3NET  | 0.8353 | 0.8385 | 0.0631 | 0.8603 | 0.8615 | 0.8728 | 0.0296 | 0.9117 |
| ICNET  | 0.8539 | 0.8346 | 0.0695 | 0.8538 | 0.8700 | 0.8843 | 0.0281 | 0.9226 |
| DCMF   | 0.8193 | 0.8199 | 0.0677 | 0.8588 | 0.8837 | 0.8549 | 0.0348 | 0.9004 |
| Ours   | 0.8541 | 0.8508 | 0.0614 | 0.8671 | 0.8699 | 0.8839 | 0.0274 | 0.9119 |

| Models | DUT-RGBD | STERE |
|--------|----------|-------|
|        | F        | avgF  | MAE | S   | F   | avgF | MAE | S   |
| FCPFP  | 0.8022  | 0.7859 | 0.0754 | 0.8079 | 0.8301 | 0.8414 | 0.0513 | 0.8792 |
| CMW    | 0.8660  | 0.8683 | 0.0616 | 0.8668 | 0.8690 | 0.8724 | 0.0433 | 0.9050 |
| D3NET  | 0.7588  | 0.7221 | 0.1008 | 0.7593 | 0.8592 | 0.8661 | 0.0458 | 0.8986 |
| ICNET  | 0.8365  | 0.8306 | 0.076  | 0.8346 | 0.8647 | 0.8695 | 0.0446 | 0.9025 |
| DCMF   | 0.756   | 0.747  | 0.1028 | 0.7845 | 0.8406 | 0.8429 | 0.0543 | 0.8829 |
| Ours   | 0.8681  | 0.8712 | 0.0523 | 0.8456 | 0.8653 | 0.8790 | 0.0420 | 0.9041 |

| Models | RGBD135 |
|--------|---------|
|        | F        | avgF  | MAE | S   |
| CPFP   | 0.8187  | 0.8278 | 0.0405 | 0.8731 |
| CMW    | 0.8903  | 0.9143 | 0.0294 | 0.9325 |
| D3NET  | 0.8702  | 0.8613 | 0.0311 | 0.8977 |
| ICNET  | 0.899   | 0.8935 | 0.0266 | 0.92  |
| DCMF   | 0.82    | 0.8226 | 0.0401 | 0.8767 |
| Ours   | 0.90    | 0.8936 | 0.0259 | 0.906 |
Through the above results, it can be found that improving the quality of the depth map helps to improve the accuracy of detection, which is more advantageous than other models.

4. Conclusion
In this article, we propose a RGBD saliency detection model based on depth enhancement, which is mainly composed of the depth enhancement module (DEM), which realizes the improvement of the accuracy of saliency detection by processing the quality of the depth map. Experiments show that on five challenging data sets, our model has superior performance on four evaluation indicators.

References
[1] Han, J., Chen, H., Liu, N., Yan, C., & Li, X. (2017). CNNs-based RGB-D saliency detection via cross-view transfer and multiview fusion. IEEE transactions on cybernetics, 48(11), 3171-3183.
[2] Zhao, J. X., Cao, Y., Fan, D. P., Cheng, M. M., Li, X. Y., & Zhang, L. (2019). Contrast prior and fluid pyramid integration for RGBD salient object detection. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 3927-3936).
[3] Fan, D. P., Zhai, Y., Borji, A., Yang, J., & Shao, L. (2020, August). BBS-Net: RGB-D salient object detection with a bifurcated backbone strategy network. In European Conference on Computer Vision (pp. 275-292). Springer, Cham.
[4] Fu, K., Fan, D. P., Ji, G. P., & Zhao, Q. (2020). JL-DCF: Joint learning and densely-cooperative fusion framework for RGB-D salient object detection. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 3052-3062).
[5] Pang, Y., Zhang, L., Zhao, X., & Lu, H. (2020). Hierarchical dynamic filtering network for rgb-d salient object detection. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXV 16 (pp. 235-252). Springer International Publishing.
[6] Zhang, M., Fei, S. X., Liu, J., Xu, S., Piao, Y., & Lu, H. (2020, August). Asymmetric two-stream architecture for accurate RGB-D saliency detection. In European Conference on Computer Vision pp. 374-390.

[7] 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, 32(5), 2075-2089.

[8] Zhang, J., Fan, D. P., Dai, Y., Anwar, S., Saleh, F. S., Zhang, T., & Barnes, N. (2020). Uc-net: Uncertainty inspired rgb-d saliency detection via conditional variational autoencoders. In Proceedings of the IEEE conference on computer vision and pattern recognition .pp. 8582-8591.

[9] Li, G., Liu, Z., Ye, L., Wang, Y., & Ling, H. (2020). Cross-modal weighting network for RGB-D salient object detection. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVII 16.pp. 665-681.

[10] Li, G., Liu, Z., & Ling, H. (2020). ICNet: Information conversion network for RGB-D based salient object detection. IEEE Transactions on Image Processing, 29, 4873-4884.

[11] Chen, H., Deng, Y., Li, Y., Hung, T. Y., & Lin, G. (2020). RGBD salient object detection via disentangled cross-modal fusion. IEEE Transactions on Image Processing, 29, 8407-8416.