Multi-level interactions for RGB-D object detection

Qi Li *, Meng Dai and Qing Zhang

School of Computer Science and Information Engineering, Shanghai Institute of Technology, Shanghai, 201418, China

*Corresponding author’s e-mail: 206141141@mail.sit.edu.cn

Abstract. In order to efficiently utilize high-level information and depth information in RGB-D saliency object detection, multi-level fusion is studied. Different from existing methods which ignore feature dilution in the process of feature downward transmission, a multi-level interactive fusion method is designed and compared with five advanced models through four evaluation indexes. The experimental results show that the model in this paper is advanced.

1. Introduction

Saliency object detection mainly uses computer to imitate human vision to detect or extract the most interesting object or a certain area in the image. At present, saliency object detection has been widely used in target segmentation and recognition, image editing, visual tracking and other computer fields. However, when performing the saliency object detection of a complex scene (the foreground and the background are extremely similar, and there are multiple objects in the image), a model based on the input of the red, green and blue three-channel color map (RGB) alone cannot get a very good result. One of the main reasons for this is the lack of depth information that can reflect the spatial structure. With the emergence of low-cost but high-performance depth sensors, researchers began to combine depth maps (Depth) for saliency detection. The pixels of the depth map reflect the distance of the object from the sensor. This method of combining the depth map is called RGB-D significance detection.

In order to effectively fuse RGB and Depth graphs across modes, Chen [1] proposed an RGB-D significance detection algorithm for multi-scale residual rough prediction. Feng [2] extracted RGB features at each stage and then combined them with depth features. Ji [3] proposed a cooperative learning framework for computing significance detection.

Although the above methods have a good effect in significance detection, there are still some deficiencies. First, they do not take into account the difference between RGB and Depth modes, just by simple addition or channel series. In addition, most of the top-down cross-modal fusion methods ignore the phenomenon of feature dilution in the continuous downward transmission of high-level features, which will affect the effect of feature fusion.

To solve the above problems, this paper proposes a new mode fusion module based on attention mechanism, which is embedded into the whole coding-decoding network structure, and simultaneously carries out multiple downward transmission of high-level features. Compared with eight other recently published advanced models, our model achieves better detection performance on five widely used public datasets in multiple evaluation indicators.

2. The proposed method

The model structure proposed in this paper is the coding-decoding dual-flow structure as shown in Figure 1. The encoder is composed of two VGG-16 feature extraction networks for feature extraction
across modes. We use hierarchical fusion architecture to fuse multi-scale features and perform side output estimation. In this part, the specific operation of the fusion method used in this paper is given.

2.1. Network structure
As shown in Figure 1, the RGB and Depth plots are individually fed into two identical trunk branch networks. In order to improve the computational efficiency of the model, VGG-16 with relatively shallow layers is used as the backbone branch network to extract relevant features. The whole VGG-16 trunk branch network is divided into five blocks. In this paper, the features extracted from the last convolutional layer of each block are taken for cross-modal feature fusion operation. RGB streams are used to extract the main feature information of an image, such as color, position or other low-level features, as well as other high-level semantic information and contextual features. The Depth stream mainly captures spatial information to make saliency detection more accurate and complete. In order to better integrate the two, a fusion module is designed in the fusion stage. At the same time, high-level features will be transmitted to each underlying fusion module to eliminate the dilution problem in the process of feature transmission, and each result of the trunk branch will be decoded side output.

2.2. Implementation details
Due to the attribute difference of different modes, the complementary information of RGB and Depth can not be effectively mined and fully used simply by adding or connecting the equal weight values of RGB and Depth. In order to improve the compatibility and merging of RGB and Depth, and realize the filtering of unnecessary information, this paper uses the attention mechanism to extract more effective features in the fusion module.
First, RGB and Depth as well as the advanced features obtained by feedback are added at the pixel level. Then, continuous attention mechanism operation is performed, and finally, a convolution is performed to extract more features.

\[ f_{\text{rgb}} \cdot f_{\text{rgb}} \] represents the result generated by the last convolution layer of each block, and \( n \) represents the number of layers.

\[ f_{\text{CFM}}^i = f_{\text{rgb}}^i \oplus f_{\text{depth}}^i \oplus S_{\text{CFM}}^i \] (1)

\[ S_{\text{CFM}}^i = \text{Conv}(\text{SA} \{ \text{CA}(f_{\text{CFM}}^i) \}) \] (2)

If \( S_{\text{CFM}}^i \) represents the pooled and pixel-level summation results of all feature results higher than the current layer \( i \), \( f_{\text{CFM}}^i \) represents the result of a transition operation in a fusion module, \( S_{\text{CFM}}^1 \) represents the prediction result of current layer \( 1 \).

2.3. Training

In the training stage, we supervise each side output, and the loss function of the whole model is also composed of the loss function of each side output.

\[ \text{LOSS} = y \log x + (1 - y) \log(1 - x) + (A \cap B)(A \cup B)^{-1} \] (3)

\( y \) represents the category, \( x \) represents prediction category, \( A \) represents the prediction area, \( B \) represents the real area.

Use the same data as [4] for training. The size of each image input is 352*352 and the batch size is 4.

3. The experimental results

In order to prove the reliability of the whole model, this paper refers to [4][5][6][7] and uses five data sets, namely NLPR, SIP, DUT-RGBD, STERE and RGBD135. In this paper, the model is compared with advanced models at the present stage that are also based on RGB-D for significance detection, including CPFP[7], CMW[8], D3NET[5], ICNET[9], DCMF[10]. At the same time, four evaluation indicators were used, and they were S-measure, F-measure, MAE, avgF.

S-measure mainly evaluates the structural similarity between significance plots and binary truth plots, and its formula is as follows:

\[ S = \alpha S_o + (1 - \alpha)S_r \] (4)

\( S_o \) represents object perception, \( S_r \) represents region perception, and \( \alpha \) is an equilibrium parameter, generally set to 0.5. And F-measure mainly calculates the weighted average between the accuracy and recall rate of the binarization significance graph, which is calculated as follows:

\[ F = (\beta^2 + 1)P(R(\beta^2P + R)^{-1}) \] (5)

\( \beta^2 \) is a hyperparameter used to assign different weights to accuracy and recall. avgF is the average of f-measure. MAE evaluates the mean absolute error of all pixels between the saliency graph and truth graph. MAE is calculated as follows:

\[ MAE = |S_i - g_i|T^{-1} \] (6)
$T$ represents the total number of pixels, $s_i$ represents a salient graph pixel, $g_i$ represents truth graph pixels.

3.1. Table and visual results
This paper compares the detection results with other advanced models and presents the results more intuitively by means of tables and visual images.

| 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 | 0.8387 | 0.8549 | 0.0348 | 0.9004 |
| Ours   | 0.8522 | 0.8509 | 0.0603 | 0.8665 | 0.8788 | 0.8819 | 0.0268 | 0.9201 |

| 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   | 0.8682 | 0.8721 | 0.0453 | 0.8468 | 0.8657 | 0.8790 | 0.0418 | 0.9035 |

| 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   | 0.9001 | 0.8937 | 0.0258 | 0.9071 |
Through the above data and pictures, it can be found that this paper not only maintains advantages over other methods in terms of various data, but also has advantages in the detection results in complex scenes.

4. Conclusion
In order to solve the problem of feature dilution in the top-down process, the high-level information is fed back to the low-level features for many times to achieve the optimal fusion of the two. And attention mechanism is used to extract information in the process of fusion to realize the interaction between high and low levels of information. Through experimental comparison, the superiority of the proposed method is proved effectively.

References
[1] Chen S., Fu Y. (2020) Progressively Guided Alternate Refinement Network for RGB-D Salient Object Detection. In: Vedaldi A., Bischof H., Brox T., Frahm JM. (eds) Computer Vision – ECCV 2020. pp.520-538
[2] Feng, D., Barnes, N., You, S., & McCarthy, C. (2016). Local background enclosure for RGB-D salient object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2343-2350.
[3] Ji, W., Li, J., Zhang, M., Piao, Y., & Lu, H. (2020). Accurate rgb-d salient object detection via collaborative learning. In Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16. pp. 52-69.
[4] 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.
[5] 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.
[6] 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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.