Depth Guided Cross-modal Residual Adaptive Network for RGB-D Salient Object Detection

Zhengyun Zhao\(^1,\)\(^a\), Qingpeng Yang\(^1,\)\(^b\), Shangqin Yang\(^1,\)\(^c\) and Jun Wang\(^1,\)\(^2,\)\(^d\)\(^*\)

\(^1\)School of Artificial Intelligence, Henan University, Kaifeng, Henan, 475004, China
\(^2\)Henan Key Laboratory of Big Data Analysis and Processing, Henan University, Kaifeng, Henan, 475004, China
\(^a\)email: zhaozhengyun@henu.edu.cn, \(^b\)email: yqp@henu.edu.cn, \(^c\)email: 565361370@qq.com
\(^*\)Corresponding author: \(^d\)email: auwangjun@henu.edu.cn

Abstract. Depth modal features can provide complementary information for salient object detection (SOD). Most of the existing RGB-D SOD methods focus on fully combining RGB and Depth modal features without distinguishing them. In this paper, we propose a new depth guided cross-modal residual adaptive network for RGB-D SOD. We use two independent resnet-50 to extract the features of the two modes respectively. Then the cross-modal channel-wise refinement module is designed to obtain complementary modal information. We design a cross-modal guided module to make complementary modal information guide RGB image feature extraction. Finally, the residual adaptive selection module is used to enhance the spatial mutual concerns between the two modal features to achieve multimodal information fusion. Experimental results show that our method can achieve a more reasonable fusion state of RGB and Depth, and verify the effectiveness of our final saliency model.

1. Introduction
Salient object detection (SOD) aims to separate the most salient object from the background. The SOD has been applied in various computer vision tasks, such as image understanding[1], co-saliency detection[2], semantic segmentation[3]. In recent years, the Depth information of images makes the RGB-D SOD widely studied. People improve the performance of saliency detection by the complementary different features of RGB and Depth modes.

The traditional RGB-D SOD methods adopt the early fusion strategy[4] to combine RGB appearance information and Depth cues in the images. However, there are great differences between the two modes, so it is difficult to integrate them well by traditional methods. With the rise of the convolutional neural network, more and more people apply deep learning technology to RGB-D SOD and achieve good results. Han et al.[5]used a joint representation layer to fuse RGB and Depth modes representation layers. Chen et al.[6]proposed a complementary perceptual fusion module to capture cross-modal and cross-level features. In this paper, we design a new depth guided cross-modal residual adaptive network for RGB-D SOD. Through the design of cross-modal fusion and cross-modal guidance modules, our network focuses more on the fusion of multimodal attention. Besides, to enhance the spatial mutual concerns between the two modal features, the residual adaptive selection module is designed.

The main works of this paper are as follows:
1. A new depth guided cross-modal residual adaptive network for RGB-D Salient object detection is
proposed. It can effectively complement RGB and Depth modes and guide image feature extraction.

2. In the process of RGB and Depth modes fusion, cross-modal channel-wise refinement (CMCR) module and cross-modal guided (CMG) module are designed for information guidance. Then, the residual adaptive selection (RAS) module is designed to enhance the mutual concerns between features.

3. We test the proposed method on four public datasets, and the experimental results show the effectiveness of the proposed model.

2. Methodology

In this section, we first describe the overall network structure, then introduce the key components of the network structure, including cross-modal channel-wise refinement (CMCR) module in Subsection 2.2, cross-modal guided (CMG) module in Subsection 2.3, residual adaptive selection (RAS) module in Subsection 2.4, and loss function in Subsection 2.5.

2.1. Model Architecture

Figure 1. shows the proposed depth guided cross-modal residual adaptive network (DCRANet) architecture. First, two identical resnet-50[9] are used as network encoders. Then, we design CMCR module which sends the fused cross-modal information to each layer of RGB feature extraction step by step through CMG module. In the decoding process, we design RAS module to select the weight of mutual attention, and abandon the useless background information.

2.2. Cross-modal fusion strategy

We design an effective cross-modal fusion strategy CMCR module to fully extract and fuse the cross-modal features. As shown in Figure 2, the top-level features R5 and D5 are used as input features of the CMCR module. After a simple weight layer coding operation, the common pixels in the features are enhanced and the blurred pixels are reduced by the multiplication of image pixels. In learn the input feature residuals, we add the encoded output and the multiplication output to the image pixels. The process is as follows:

\[ f^c = \delta_1(R5) \oplus \delta_2(D5) \oplus (\delta_1(R5) \otimes \delta_2(D5)) \]  

(1)

Where \( \oplus \) and \( \otimes \) denote element-wise summary multiplication operation perspective. Each encoder \( \delta_{1,2} \) is typically composed of a 3 × 3 volatile layer followed by BN and Relu.

After pixel multiplication and pixel addition operations, rich feature \( f^c \) is obtained. To highly respond to strong object features in the channel, we use global features to understand the context of
attention weight. First, $f^c$ is compressed by a global average pooling, followed by full join and Relu operations, and finally mapped to $[0,1]$ by a sigmoid normalization to align the size with the feature channel. To preserve $f^c$ feature, we multiply $f^c$ feature with normalized feature by residual join to get feature $U_s$. Finally, we concatenate feature $U_s$ with input feature, and through weight layer operation again, we get output feature RD which is the same as R5 and D5 channel number. The process is as follows:

$$
g = \frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} f^c(i, j) \quad (2)$$

$$\quad U_s = (\sigma(\delta(\omega g))) \otimes f^c \quad (3)$$

$$\quad RD = U_s \otimes \delta_s (R5) \otimes \delta_s (D5) \quad (4)$$

Where $W$, $H$ and are the length and width of the feature graph respectively, and $f^c(i, j)$ is the value of each grid point in the feature graph $f^c(i, j)$. $\omega$ is the weight of full connection layer, $\delta$ is Relu activation function, $\sigma$ is sigmoid operation.

### 2.3. Cross-modal guided module

Four CMG modules with the same structure are designed to guide resnet-50 step by step to extract the feature of RGB input image. Take the CMG on the rightmost end of the network as an example, as shown in figure 2. The fourth layer feature $D_4$ in Depth feature extraction process and $RD$ are used as the input of CMG module. The output features of CMG are obtained each time to guide RGB image feature extraction and feedback to the next CMG module as input.

![Figure 2. Cross-modal guided module.](image)

The whole CMG module is divided into five branches. Four branches use four different dilation rate $r$ ($r = 1, 2, 4, 8$) of the dilation convolution layer $\{L_{r_i}\}_{i=1}^4$ to obtain more information of the receptive field, and four output features $\{L_{r_i}\}_{i=1}^4$ are obtained. The process is as follows:

$$Y_i = L_{r_i}(\text{conv}_1(up(RD) \otimes D4)) \quad (5)$$

Where up is the bilinear interpolation up-sampling operation, $L_r$ is the $3 \times 3$ convolution operation with different expansion rates, and $\text{conv}_1$ is the $1 \times 1$ convolution operation.

In the other branches, channel attention is used to weight feature $P$ and output feature $U_c$ is obtained. The channel number of the final output feature $f_p$ is consistent with that of the input feature $P$. The process is as follows:

$$U_c = \sigma(\delta(\omega(\frac{1}{W \times H} \sum_{i=1}^{W} \sum_{j=1}^{H} p(i, j)))) \otimes p \quad (6)$$

$$f_p = \text{conv}_1(P \otimes \{Y_i\}_{i=1}^4) \quad (7)$$

Where $\omega$ is the weight of the full connection layer, $\delta$ is the Relu and $\sigma$ is the sigmoid.
2.4. Residual adaptive selection module

We design five of the same structure RAS module. Take the RAS module on the rightmost end of the network specifically, as shown in figure 3. the top-level R5 of RGB image feature extraction and the feature RD obtained by CMCR module are the input features of RAS module. Firstly, we adopt convolution layers $1 \times k$ and $k \times 1$ to capture high-resolution spatial concerns, which can reduce the operation parameters and obtain multi-scale information. The process is as follows:

\[
X_1 = \text{concat}(\text{conv}_1(\text{conv}_2(\text{RD})), \text{conv}_2(\text{conv}_3(\text{RD})))
\]

\[
X_2 = \text{concat}(\text{conv}_1(\text{conv}_2(\text{R5})), \text{conv}_2(\text{conv}_3(\text{R5})))
\]

On the other hand, we deal with RD and R5 by concatenation, sigmoid and $1 \times 1$ convolution, and multiply their output features with $X_1$ and $X_2$ elements after sigmoid. After obtaining two sets of selective adaptive features, they are added to the $X_1$ and $X_2$ elements after sigmoid respectively. Finally, we connect these features through residual connection and get the output feature $S_1$. The operation process is as follows:

\[
Y_1 = \text{conv}_1((\text{conv}_1(\sigma(\text{RD} \odot \text{R5})) \otimes \sigma(X_1)) \oplus \sigma(X_1))
\]

\[
Y_2 = \text{conv}_1((\text{conv}_1(\sigma(\text{RD} \odot \text{R5})) \otimes \sigma(X_2)) \oplus \sigma(X_2))
\]

\[
S_1 = \text{Trans}((Y_1 \oplus Y_2) \odot X_1 \odot X_2)
\]

Where Trans means $\text{conv}1 \times 1$, batch normal and Relu activate a series of operations.

2.5. Loss function

We use the binary cross-entropy loss to optimize our network. The loss function is as follows:

\[
L_{ce}(U,G) = G \log U + (1 - G) \log (1 - U)
\]

Where $U \in [0,1]^{H \times W \times 1}$ represents the final output prediction graph of the network, and $G \in [0,1]^{H \times W \times 1}$ represents the binary ground truth salience map.

3. Experiment and result analysis

In this section, we introduce the experiment and analysis in detail. Firstly, we introduce the implementation of the experiment, then introduce the datasets and evaluation criteria. Finally, the experimental results are analyzed and compared with the advanced methods.

3.1. Implementation details

We train our model on the NJU2K[10] and NLPR[11] datasets with 2185 images. Our model is implemented in the Pytorch framework. We use stochastic gradient descent (SGD) to optimize the whole network. The batch size is 32, the momentum is 0.9, and the weight attenuation is 5e-4. The maximum learning rate is 0.01, the minimum learning rate is 1e-4, and finish training after 200 epochs.
3.2. Datasets and Evaluation criteria

Datasets. We test our model on four public datasets: NJU2K[10], NLPR[11], SSD[12] and SIP[13]. NJU2K is a large dataset with 1985 samples. NLPR contains 1000 pictures of indoor and outdoor scenes. SSD dataset is 80 high-quality images selected from indoor and outdoor scenes with a resolution of 960 x 1080. SIP dataset uses mobile phones to take data images, including 1000 real-world images of people.

Evaluation criteria. To quantitatively evaluate the effectiveness of our proposed method, we use three widely used indicators: precision recall (PR) curve, F-measure and MAE.

3.3. Results analysis

In this subsection, we compare our proposed model with six existing models, including two traditional methods LBE[7], CDB[14] and four deep learning methods DF[8], PCF[6], DMRA[15], MCINet[16].

Visual comparisons. To show the advantages of our method more intuitively, we select some pictures from different datasets and compare them with the other six methods, as shown in figure 4. Compared with other methods, our method can highlight the salient object clearly in various challenging situations.

PR curve. We compare our method with six existing algorithms using PR curve. As shown in figure 5. We depict the PR curves generated by our method and existing methods on NLPR and SSD datasets. It is easy to infer from the PR curve that our method (red solid line) is superior to all other algorithms.

F-measure and MAE. As shown in table 1. we compare our model with the other six models in F-measure and MAE scoring. From the data, we can see that our method has absolute advantages.

| model | LBE | CDB | DF | PCF | DMRA | MCINet | Ours |
|-------|-----|-----|----|-----|------|--------|------|
| NJU2K | Fβ↑ | 0.606 | 0.498 | 0.653 | 0.840 | 0.873  | 0.876 | 0.889 |

Figure 4. Visual comparisons of the proposed method and the other algorithms.

Figure 5. PR Curves of the proposed model and other algorithms.
4. Conclusions
In this paper, we consider the Depth information in the image and design a new salient object detection network. The CMCR module is used to learn the weight coefficients of each channel in the RGB and Depth modes, which enhances the discrimination ability of the model for each channel feature. Then, the CMG module is adopted to guide the network to extract the features. Then, the RAS module is used to enhance the spatial mutual attention between the two modes. Finally, the binary cross-entropy loss function is used to guide the network to detect salient objects. Experimental results on public datasets demonstrate the effectiveness of the proposed model components and our final saliency model.

Acknowledgments
This work was supported by the Henan Provincial Department of Science and Technology Research Project (No. 212102210156).

References
[1] Zhu, J. Y., Wu, J., Xu, Y., Chang, E., & Tu, Z. (2014). Unsupervised object class discovery via saliency-guided multiple class learning. IEEE transactions on pattern analysis and machine intelligence, 37(4): 862-875.
[2] Fan, D. P., Li, T., Lin, Z., Ji, G. P., Zhang, D., Cheng, M. M., Fu, H., & Shen, J. (2020). Re-thinking cosalient object detection. arXiv preprint arXiv:2007.03380, 2020.
[3] Shimoda, W., & Yanai, K. (2016) Distinct class-specific saliency maps for weakly supervised semantic segmentation. In: European Conference on Computer Vision. Amsterdam. pp. 218-234.
[4] Song, H., Liu, Z., Du, H., Sun, G., Le Meur, O., & Ren, T. (2017). Depth-aware salient object detection and segmentation via multiscale discriminative saliency fusion and bootstrap learning. IEEE Transactions on Image Processing, 26(9), 4204-4216.
[5] Han, J., Chen, H., Liu, N., Yan, C., & Li, X. (2017). CNNs-based RGB-D saliency detection and segmentation via cross-view transfer and multiview fusion. IEEE Transactions on Cybernetics, 48(11), 3171-3183.
[6] 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.
[7] Feng, D., Barnes, N., You, S., & McCarthy, C. (2016). Local background enclosure for RGB-D salient object detection. In: Computer Vision and Pattern Recognition. Las Vegas. pp. 2343-2350.
[8] 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.
[9] He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In: computer Vision and Pattern Recognition. Las Vegas. pp. 770-778.
[10] 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.
[11] Peng, H., Li, B., Xiong, W., Hu, W., & Ji, R. (2014, September). Rgbd salient object detection: a benchmark and algorithms. In: European Conference on Computer Vision. Amsterdam. pp. 92-109.
[12] Zhu, C., & Li, G. (2017). A three-pathway psychobiological framework of salient object detection using stereoscopic technology. In: International Conference on Computer Vision. Venice. pp. 3008-3014.

[13] 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. DOI:10.1109/TNNLS.2020.2996406.

[14] Liang, F., Duan, L., Ma, W., Qiao, Y., Cai, Z., & Qing, L. (2018). Stereoscopic saliency model using contrast and depth-guided-background prior. Neurocomputing, 275: 2227-2238.

[15] Piao, Y., Ji, W., Li, J., Zhang, M., & Lu, H. (2019). Depth-induced multi-scale recurrent attention network for saliency detection. In: International Conference on Computer Vision. Seoul. pp. 7254-7263.

[16] Huang, Z., Chen, H. X., Zhou, T., Yang, Y. Z., & Wang, C. Y. (2020). Multi-level Cross-modal Interaction Network for RGB-D Salient Object Detection. arXiv preprint arXiv:2007.14352.