M²RNet: Multi-modal and Multi-scale Refined Network for RGB-D Salient Object Detection

Xian Fang, Jinchao Zhu, Ruixun Zhang, Xiuli Shao, Hongpeng Wang

Abstract

Salient object detection is a fundamental topic in computer vision. Previous methods based on RGB-D often suffer from the incompatibility of multi-modal feature fusion and the insufficiency of multi-scale feature aggregation. To tackle these two dilemmas, we propose a novel multi-modal and multi-scale refined network (M²RNet). Three essential components are presented in this network. The nested dual attention module (NDAM) explicitly exploits the combined features of RGB and depth flows. The adjacent interactive aggregation module (AIAM) gradually integrates the neighbor features of high, middle and low levels. The joint hybrid optimization loss (JHOL) makes the predictions have a prominent outline. Extensive experiments demonstrate that our method outperforms other state-of-the-art approaches.

Keywords: Saliency detection, Deep learning, Multi-modal feature, Multi-scale feature, Loss function

1. Introduction

Salient object detection (SOD) aims to identify the most conspicuous object that attracts humans in the scene. It has been successfully applied in various fields, such as image retrieval, robot navigation, person re-identification and many more.

The SOD methods have exhibited broad prospects owing to the powerful representation ability of convolutional neural networks (CNNs) and fully convolutional networks (FCNs). Most of them resort to a single RGB information, which is difficult to achieve satisfactory results in complex scenes. At present, depth information has become growing popular thanks to the emergence of affordable and portable devices. As a supplement to RGB features, depth features provide rich distance information. However, the inherent differences between the multiple modalities lead to the bottleneck of feature fusion. Moreover, although the features of each scale have the detail or semantic information, it is hard to adequately aggregate them.

To this end, we propose a novel multi-modal and multi-scale refined network (M²RNet) for RGB-D salient object detection. Specifically, the network is composed of the presented nested dual attention module (NDAM), adjacent interactive aggregation module (AIAM) and joint hybrid optimization loss (JHOL). In NDAM, we sequentially leverage channel and spatial attention to explicitly understand what and where is meaningful, thereby emphasizing or suppressing RGB and depth information that is important or unnecessary, for the purpose of boosting the fusion of RGB and depth features. In AIAM, we leverage the interaction of progressive and jumping connections in parallel to gradually learn information in abundant resolution for the purpose of boosting the aggregation of high level, middle level and low level features. In JHOL, we are devoted to guaranteeing inter-class discrimination and intra-class consistency by taking the local and global correlation of each pixel into account. These three components work together to achieve remarkable detection results. The visual comparison is shown in Fig. 1. Comparing with the saliency maps obtained by state-of-the-art approaches on several examples, those of our method are more exact.

Our main contributions are summarized as follows:

- We propose a multi-modal and multi-scale refined network (M²RNet), which is equipped with the presented nested dual attention module (NDAM), adjacent interactive aggregation module (AIAM) and joint hybrid optimization loss (JHOL) components. Our network is capable of re-
fining the multi-modal and multi-scale features simultaneously, nearly without extra computing cost under specific supervision.

- We conduct extensive experiments on seven datasets and demonstrate that our method achieves consistently superior performance against 12 state-of-the-art approaches including three RGB salient object detection approaches and nine RGB-D salient object detection approaches in terms of six evaluation metrics.

2. Related Work

2.1. RGB Saliency Detection

Lots of RGB salient object detection methods have been developed during the past decades.

For instance, Zhang et al. [10] proposed a PAGR, which selectively integrates multi-level contextual information. Liu et al. [11] proposed a PiCANet, which generates attention over the context regions for each pixel. Su et al. [12] proposed a BANet, which enhances the feature selectivity at boundaries and keeps context regions for each pixel. Zhao et al. [13] proposed an EGNet, which explores edge information to preserve salient object boundaries. Wu et al. [14] proposed a CPD, which uses cascaded partial decoder to discards low-level features. Liu et al. [15] proposed a PoolNet, which explores the potentials of pooling. Zhang et al. [16] proposed a CapSal, which uses image captioning for detecting. Qin et al. [17] proposed a BAS-Net, which focuses on end-to-end boundary-aware. Pang et al. [18] proposed a MINet, which exchanges information between multi-scale. Wei et al. [19] proposed a LDF, which decouples the saliency label into body map and detail map for iterative information exchange.

2.2. RGB-D Saliency Detection

Most recently, RGB-D salient object detection methods have rapidly aroused the concern of researchers and made impressive progress.

For instance, Chen et al. [20] proposed a TANet, which combines the bottom-up stream and the top-down stream to learn cross-modal complementarity. Wang et al. [21] proposed an AFNet, which adaptively fuses the predictions from the separate RGB and depth streams using the switch map. Piao et al. [22] proposed a DMRa, which involves residual connections, multi-scale weighting and recurrent attention. Zhao et al. [8] proposed a CPFP by making use of feature fusion of contrast prior and fluid pyramid. Fan et al. [23] proposed a D³Net, which automatically discards the low-quality depth maps via

Figure 2: The overall architecture of the proposed M²RNet. The backbone of our network is VGG-16 [9].
gate connection. Li et al. [24] proposed an ICNet, which can learn the optimal conversion of RGB features and depth features to autonomously merge them. Pang et al. [25] proposed a HDFNet, in which the features of the network are densely connected and through the dynamic expansion pyramid. Fan et al. [26] proposed a BBS-Net, in which multi-level features are partitioned into teacher and student features in the cascade network. Zhao et al. [27] proposed a DANet, which explores early fusion and middle fusion between RGB and depth. Zhao et al. [28] proposed a CMWNet, which weights the fusion of low, medium and high levels to encourage feature interaction. Fu et al. [29] proposed a JL-DCF for joint learning and densely-cooperative fusion. Zhang et al. [30] proposed an A2dele, which learns the distribution of saliency maps by conditional variational autoencoder. Piao et al. [31] proposed an A2dele, which uses network prediction and attention as two bridges to transfer deep knowledge from deep stream to RGB stream. Liu et al. [26] proposed a BBS-Net, in which multi-level features are connected and through the dynamic expansion pyramid.

3. Proposed Method

3.1. Network Overview

The proposed multi-modal and multi-scale refined network (M2RNet) is an encoder-decoder architecture, covering nested dual attention module (NDAM), adjacent interactive aggregation module (AIAM) and joint hybrid optimization loss (JHOL), as shown in Fig. 2. To be concise, we denote the output features of RGB branch in the encoder as \( f_{rgb}^1(i = 1, 2, 3, 4, 5) \), the output features of depth branch in the encoder as \( f_{rgb}^i(i = 1, 2, 3, 4, 5) \), and the output features in the decoder as \( f_{rgb}^i(i = 1, 2, 3, 4, 5) \). Let \( f_{cm}^i(i = 3, 4, 5) \) denote the combined features of RGB flow and depth flow, we utilize NDAM to strengthen their robustness to get the corresponding enhanced features \( f_{rgb}^i(i = 2, 3, 4, 5) \), where \( f_{cm}^i \) are eventually obtained after the reinforcement of nested attention. The procedure of RGB and depth feature fusion can be described as:

\[
\begin{align*}
    f_{cm}^i &= f_{rgb}^i + f_{d}, \\
    f_{cm}^5 &= S_2(C_2(S_1(C_1(f_{cm}^1)))).
\end{align*}
\]

where \( C(i) \) and \( S(i) \) denote the channel attention and spatial attention, respectively. The nested attention is roughly divided into two phases of \( P_1 \) and \( P_2 \), as shown in Fig. 3.

In the first phase \( P_1 \), given an intermediate feature \( f \in \mathbb{R}^{C \times X \times Y} \), in which \( X = H \times W \), and \( C, H \) and \( W \) are the channel, height and width of the feature \( f \), respectively, the dual attentions \( C_1(\cdot) \) and \( S_1(\cdot) \) are defined as:

\[
\begin{align*}
    C_1(f) &= \delta(Conv(f) \times Conv(f)^T \times Conv(f) + f), \\
    S_1(f) &= Conv(f) \times \delta(Conv(f)^T \times Conv(f)) + f.
\end{align*}
\]

where \((\cdot)^T\) denotes the transpose operation and \( \delta(\cdot) \) is the Softmax function. Note that the channel of the feature in Eq. (1) is set to 1/8 of the original channel of that for computation efficiency.

In the second phase \( P_2 \), for the feature \( f \in \mathbb{R}^{C \times H \times W} \), which can be reshaped with the feature \( f \), the dual attentions \( C_2(\cdot) \) and \( S_2(\cdot) \) are defined as:

\[
\begin{align*}
    C_2(f) &= \sigma(MLP(GMP(f))) \circ f, \\
    S_2(f) &= \sigma(Conv(GMPC(f))) \circ f,
\end{align*}
\]

where \( \circ \) denotes the element-wise multiplication, \( \sigma(\cdot) \) is the Sigmoid function. \( MLP(\cdot) \) represents the multi-layer perceptron, \( GMP(\cdot) \) and \( GMPC(\cdot) \) represent the global max pooling.
operation and global max pooling along the channel operation, respectively. Note that we use a global max pooling rather than a global average pooling since our goal is to find the area with the biggest visual influence.

3.3. Adjacent Interactive Aggregation Module

In general, shallower features have more detail information, while deeper features have more semantic information. The aggregation of multi-level features with different resolutions enables the context information to be integrated as sufficient as possible.

Relying on this, we propose an adjacent interactive aggregation module (AIAM) to guide the interaction of multi-scale features. In this way, the neighbor features that are of intimate correlation are constantly complemented. The interaction behind the features of three triples \( \{ f_{i}^{rbg}, f_{i}, f_{i+1} \} \) \((i = 2, 3, 4)\) helps the generation of desired resulting features \( f_{i+1}^{rbg} \) \((i = 2, 3, 4)\).

The procedure of different levels feature aggregation can be described as:

\[
    f_{i+1}^{rbg} = I_{1}(f_{i}^{rbg}, f_{i}, f_{i+1}^{rbg}) + I_{2}(f_{i}^{rbg}, f_{i+1}^{rbg}),
\]

where \( I_{1} \) and \( I_{2} \) denote two kinds of feature interactions, as shown in Fig. 4. The difference between them is that the interaction of the former is progressive, while that of the latter is jumping in the initial stage. After that, these relevant features repeatedly pass through some convolutional layers, batch normalization layers and ReLU layers, and finally are associated with a residual block.
### 3.4. Joint Hybrid Optimization Loss

Let $P = \{p|0 < p < 1\}$ and $G = \{g|0 < g < 1\}$ denote the prediction saliency map and ground truth saliency map, respectively. As the most classical loss function, binary cross entropy loss (BCEL), symbolized by $L_{BCE}$, can be formulated as:

$$L_{BCE} = - \sum_{h=1}^{R} \sum_{w=1}^{W} g \log p + (1 - g) \log(1 - p),$$  \hspace{1cm} (9)
where $H$ and $W$ are the height and width of the image, respectively.

To reinforce the capability of supervision from both local and global aspects, we propose a joint hybrid optimization loss (JHOL), symbolized by $\mathcal{L}_{JHO}$, as an auxiliary loss function, which can be formulated as:

$$
\mathcal{L}_{JHO} = \lambda_1 \mathcal{L}_1 + \lambda_2 \mathcal{L}_2 + \lambda_3 \mathcal{L}_3 + \lambda_4 \mathcal{L}_4,
$$

(10)

where $\lambda_1$, $\lambda_2$, $\lambda_3$ and $\lambda_4$ are the parameters to control the trade-off among the four terms of the loss. For the sake of simplicity, they are all set to 1. To be specific, $\mathcal{L}_1$, $\mathcal{L}_2$, $\mathcal{L}_3$ and $\mathcal{L}_4$ are defined as:

$$
\mathcal{L}_1 = \frac{H \sum_{h=1}^{H} \sum_{w=1}^{W} (p(1-g))}{H \sum_{h=1}^{H} \sum_{w=1}^{W} p},
$$

(11)

$$
\mathcal{L}_2 = \frac{H \sum_{h=1}^{H} \sum_{w=1}^{W} (g(1-p))}{H \sum_{h=1}^{H} \sum_{w=1}^{W} g},
$$

(12)

$$
\mathcal{L}_3 = \frac{H \sum_{h=1}^{H} \sum_{w=1}^{W} (p(1-g) + g(1-p))}{H \sum_{h=1}^{H} \sum_{w=1}^{W} (p + g - pg)},
$$

(13)

$$
\mathcal{L}_4 = \frac{H \sum_{h=1}^{H} \sum_{w=1}^{W} ((1-p)(1-g))}{H \sum_{h=1}^{H} \sum_{w=1}^{W} (1-p(1-g) - g(1-p))}.
$$

(14)

On the one hand, $\mathcal{L}_1$ and $\mathcal{L}_2$ focus on the local correlation of each pixel to ensure a certain degree of discrimination. On the other hand, $\mathcal{L}_3$ and $\mathcal{L}_4$ focus on the global correlation of each pixel to ensure a certain degree of consistency. As a result, $\mathcal{L}_{JHO}$ jointly makes the model capture the foreground region as smoothly as possible and filter the background region as steadily as possible.

Therefore, the total loss $\mathcal{L}_{Total}$ can be written as:

$$
\mathcal{L}_{Total} = \mathcal{L}_{BCE} + \mu \mathcal{L}_{JHO},
$$

(15)

where $\mu$ is the parameter to control the trade-off between $\mathcal{L}_{BCE}$ and $\mathcal{L}_{JHO}$. In practice, it is also set to 1.

4. Experiments

4.1. Experimental Setup

Datasets. We choose seven benchmark datasets including STEREO [36], NLPR [37], RGBD135 [38], LFSD [39], NJU2K [40], DUT-RGBD [22] and SIP [23] as the experimental material. STEREO, also known as SSBI1000, contains 1000 pairs of stereoscopic images gathered from the Internet. NLPR contains 1000 images taken under different illumination conditions. R GBD135 is also called DES, which contains 135 images about some indoor scenarios. LFSD is relatively small and contains 100 images. NJU2K contains 1985 images collected from the Internet, 3D movies and photographs. DUT-RGBD contains 1200 images taken in varied real-life situations. SIP is relatively new, which contains 929 human images. Following [25], we use 700 samples from NLPR, 1485 samples from NJU2K, and 800 samples from DUT-RGBD as the training set. The remaining samples and other datasets are used as the testing set.

Evaluation Metrics. We employ six typical evaluation metrics including S-measure ($S_\alpha$) [41], maximum F-measure ($F_{\alpha\beta}^{max}$) [42], average F-measure ($F_{\alpha\beta}^{ave}$) [42], weighted F-measure ($F_{\alpha\beta}^w$) [43], E-measure ($E_\beta$) [44] and mean absolute error ($M$) [45] to comprehensively evaluate the performance of competitors. In addition, we also plot the precision-recall (PR) curves and F-measure ($F_\beta$) curves.

Implementation Details. We implement our method based on the PyTorch toolbox with a single GeForce RTX 2080 Ti GPU. The VGG-16 [9] is adopted as the backbone. For each input image, it is simply resized to 320×320 and then fed into the network to obtain prediction without any other pre-processing (e.g., HHA [46]) or post-processing (e.g., CRF [47]). To avoid over-fitting, the techniques of flipping, cropping and rotation act as data augmentation. The stochastic gradient descent (SGD) optimizer is used with the batch size of 4, the momentum of 0.9 and the weight decay of 5e-4. The whole network is stopped after 30 epochs.

4.2. Comparison with State-of-the-arts

The proposed method is compared with other 12 state-of-the-art approaches including three RGB approaches (i.e., EGGNet [13], CPD [14] and PoolNet [15]) and nine RGB-D approaches (i.e., DF [33], CTMF [34], MMCI [35], TANet [20], AFNet [21], DMRA [22], CPFP [8], D3Net [23] and S²MA [7]). For fair comparisons, all saliency maps of these methods are provided by the authors or computed by their released codes with default settings.

Quantitative Comparison. The quantitative comparison results are reported in Table 1. It can be seen that our method performs best in almost all cases. For an intuitive comparison, the PR curves and $F_\beta$ curves are shown in Fig. 5 and Fig. 6 respectively. Obviously, the curves generated by our method are closer to the top and straighter in a large range than others, which reflects its excellence and stability.

Qualitative Comparison. The qualitative comparison results are shown in Fig. 7. It can be observed that our method can handle a wide variety of challenging scenes, such as blurred foreground, cluttered background, low contrast and multiple objects. More specifically, our method yields clear foreground, clean background, complete structure and sharp boundary. These results prove that our method is able to utilize cross-modal complementary information, which can not only achieve the reinforcement from the reliable depth maps but also prevent the contamination from the unreliable depth maps.

4.3. Ablation Studies

A series of ablation studies are conducted to investigate the impact of each core component of our method. The ablation
Figure 7: Qualitative comparison of some methods on several representative examples. (a) RGB; (b) Depth; (c) Ground truth; (d) Ours; (e) D3Net [23]; (f) DMRA [22]; (g) TANet [20]; (h) DF [33].

Experiment results are reported in Table 2. The baseline, corresponding to scheme No. 1 (i.e., the 1st rows), refers to the network like FPNs [48]. It should be pointed out that the uniform evaluation metrics are redefined as a weighted sum of scores according to the proportion of each dataset in all datasets. The visual comparison results are shown in Fig. 8.

Effect of NDAM. Actually, NDAM consists of two parts, which are $P_1$ and $P_2$. By adding them into the baseline in the individual and collective manner, corresponding to scheme No. 2-3 (i.e., the 2nd and 3rd rows) and scheme No. 4 (i.e., the 4th rows), the performance is well improved. This confirms that NDAM does indeed offer additional valuable information from the channel and spatial perspectives.

Effect of AIAM. Similarly, AIAM consists of two parts, which are $I_1$ and $I_2$. By adding them into the baseline in the individual and collective manner, corresponding to scheme No. 5-6 (i.e., the 5th and 6th rows) and scheme No. 7 (i.e., the 7th rows), the performance is greatly improved. This reveals the advantage of feature aggregation of AIAM.

Effect of JHOL. Also similarly, JHOL (i.e., $L_{JHO}$) consists of four parts, which are $L_1$, $L_2$, $L_3$ and $L_4$. By adding them into the baseline in the individual and collective manner, corresponding to scheme No. 8-11 (i.e., the 8th, 9th, 10th and 11th rows) and scheme No. 12 (i.e., the 12th rows), the performance is significantly improved. In particular, when these parts are collectively added into the baseline, there are 1.4%, 3.6%, 5.6%, 2.2%, 0.8% and 1.4% improvement in terms of uniform evaluation metrics in order, respectively. This suggests that the use of JHOL is crucial for the task.

5. Conclusion

In this paper, we propose the multi-modal and multi-scale refined network named M$^2$RNet for detecting salient objects. To start with, we present the nested dual attention module, which boosts the fusion of multi-modal features via the phased channel and spatial attention. Next, we present the adjacent interactive aggregation module, which boosts the aggregation of multi-scale features in the form of progressive and jumping connections. Last but not least, we present the joint hybrid optimization loss, which alleviates the imbalance of pixels from both local and global aspects. Exhaustive experimental results demonstrate the superiority of our method over the other 12 state-of-the-art approaches.

CRediT authorship contribution statement

Xian Fang: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing - original draft, Visualization. Jinchao Zhu: Data Curation, Writing - review & editing, Visualization. Ruixun Zhang: Writing - review & editing. Xiuli Shao: Writing - review & editing. Hongpeng Wang: Writing - review & editing, Funding acquisition.
Table 2: Ablation study of our method in terms of uniform evaluation metrics. ↑ and ↓ indicate that the larger and smaller scores are better, respectively. The best results are bold.

| No. | Baseline | +NDAM | +AIAM | +JHOL | $S_α$ ↑ | $F^{max}_β$ ↑ | $F^{max}_ω$ ↑ | $F^ω_β$ ↑ | $E_ξ$ ↑ | $M$ ↓ |
|-----|----------|-------|-------|-------|---------|-------------|-------------|-------------|---------|-------|
| 1   | ✓        |       | ✓     | ✓     | 0.885   | 0.820       | 0.775       | 0.899       | 0.872   | 0.064 |
| 2   | ✓        | ✓     |       | ✓     | 0.893   | 0.833       | 0.790       | 0.907       | 0.880   | 0.060 |
| 3   | ✓        | ✓     |       | ✓     | 0.887   | 0.827       | 0.779       | 0.903       | 0.875   | 0.063 |
| 4   | ✓        | ✓     |       | ✓     | 0.894   | 0.837       | 0.796       | 0.910       | 0.882   | 0.059 |
| 5   | ✓        | ✓     | ✓     | ✓     | 0.896   | 0.839       | 0.804       | 0.914       | 0.885   | 0.058 |
| 6   | ✓        | ✓     | ✓     | ✓     | 0.890   | 0.832       | 0.788       | 0.906       | 0.878   | 0.061 |
| 7   | ✓        | ✓     | ✓     | ✓     | 0.891   | 0.865       | 0.812       | 0.919       | 0.873   | 0.055 |
| 8   | ✓        | ✓     | ✓     | ✓     | 0.891   | 0.865       | 0.812       | 0.919       | 0.873   | 0.055 |
| 9   | ✓        | ✓     | ✓     | ✓     | 0.892   | 0.812       | 0.786       | 0.899       | 0.866   | 0.061 |
| 10  | ✓        | ✓     | ✓     | ✓     | 0.893   | 0.848       | 0.819       | 0.916       | 0.876   | 0.053 |
| 11  | ✓        | ✓     | ✓     | ✓     | 0.887   | 0.819       | 0.778       | 0.900       | 0.871   | 0.063 |
| 12  | ✓        | ✓     | ✓     | ✓     | 0.899   | 0.856       | 0.831       | 0.921       | 0.880   | 0.050 |
| 13  | ✓        | ✓     | ✓     | ✓     | 0.913   | 0.873       | 0.854       | 0.929       | 0.897   | 0.043 |

Figure 8: Visual comparison of the impact of each component of our method on several examples. (a) RGB; (b) Depth; (c) Ground truth; (d) Baseline; (e) +NDAM; (f) +AIAM; (g) +JHOL; (h) +NDAM+AIAM+JHOL.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgement

This research was supported by the National Key R&D Program of China under Grant 2019YFB1311804, the National Natural Science Foundation of China under Grant 61973173, 91848108 and 91848203, and the Technology Research and Development Program of Tianjin under Grant 18ZXZNGX00340 and 20YFZCSY00830.

References

References

[1] Y. Gao, M. Shi, D. Tao, C. Xu, Database saliency for fast image retrieval, IEEE Transactions on Multimedia 17 (3) (2015) 359–369.

[2] X. Yang, X. Qian, Y. Xue, Scalable mobile image retrieval by exploring contextual saliency, IEEE Transactions on Image Processing 24 (6) (2015) 1709–1721.

[3] C. Craye, D. Filliat, J.-F. Goudou, Environment exploration for object-based visual saliency learning, in: Proceedings of the International Conference on Robotics and Automation (ICRA), 2016, pp. 2303–2309.

[4] R. Zhao, W. Ouyang, X. Wang, Unsupervised salience learning for person re-identification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3586–3593.

[5] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proceedings of the IEEE 86 (11) (1998) 2278–2324.

[6] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015, pp. 3431–3440.

[7] N. Liu, N. Zhang, J. Han, Learning selective self-mutual attention for RGB-D saliency detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 13756–13765.

[8] J. Zhao, Y. Cao, D. Fan, M. Cheng, X. Li, L. Zhang, Contrast prior and fluid pyramid integration for RGBD salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2019, pp. 3927–3936.

[9] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv preprint arXiv:1409.1556.

[10] X. Zhang, T. Wang, J. Qi, H. Lu, G. Wang, Progressive attention guided recurrent network for salient object detection, in: Proceedings of the IEEE
