SCIENCE CHINA
Information Sciences

June 2022, Vol. 65 160107:1–160107:2
https://doi.org/10.1007/s11432-020-3337-9

• LETTER •
Special Focus on Deep Learning for Computer Vision

RLLNet: a lightweight remaking learning network for saliency redetection on RGB-D images

Wujie ZHOU[1,2*], Chang LIU[1], Jingsheng LEI[1] & Lu YU[2]

[1]*School of Information and Electronic Engineering, Zhejiang University of Science and Technology, Hangzhou* 310023*, China;*
[2]*College of Information and Electronic Engineering, Zhejiang University, Hangzhou* 310023*, China*

Received 19 November 2020/Revised 28 June 2021/Accepted 24 August 2021/Published online 22 April 2022

Citation Zhou W J, Liu C, Lei J S, et al. RLLNet: a lightweight remaking learning network for saliency redetection on RGB-D images. Sci China Inf Sci, 2022, 65(6): 160107, https://doi.org/10.1007/s11432-020-3337-9

Dear editor,

In recent years, given the importance of depth information for human vision and the popularity of depth sensing, visual salient object detection (SOD) from RGB-D (color and depth) images has attracted much attention [1–3]. Most SOD methods use multimodal visual models to process stereo images and obtain complementary cues by various fusion methods to predict saliency maps, especially when dealing with complex scenes such as those having similar foreground and background. Although RGB-D SOD methods are more effective in practice than their RGB counterparts, two main limitations remain to be addressed for the segmentation of complex scenes toward more accurate SOD [4–6]. (1) Most methods focus on the extraction of salient regions using complete decoding but lack a multilevel deep network structure. (2) SOD focuses on areas of human interest and background information is equally important. To obtain results that resemble the ground truth, the longitudinal depth utilization of background and foreground information should be improved.

To address the abovementioned challenges, we propose a remaking learning lightweight framework (RLLNet) to mine hidden cues in previously generated saliency maps.

Method. The overall structure of the proposed RLLNet is shown in Figure 1. RLLNet comprises two end-to-end networks. Its hierarchical cross-modal principal network comprises the global feature extraction module (PFEM) and cross-modal interaction gate module (CIGM) to generate initial background prediction images; then, the lightweight remaking learning network composed by the global dual-attention mechanism module (GDAM) and the refinement learning module (RLM) is used to produce accurate saliency maps. We also analyze the advantages of the adopted learning approach and network supervision.

The hierarchical cross-modal principal network (including PFEM and CIGM) uses the relatively simple MobileNetV3 [7] as the basic network. We remove the last pooling layer of MobileNetV3 to better fit our visual task. The host network is responsible for generating the initial significance prediction graph and transferring the decoded information to the lightweight network. It guides the next layer of information step by step through cross-modal interactive learning combining RGB information and depth information and then generates the initial saliency prediction map with background information using the background truth.

The proposed lightweight remaking learning network is based on the MobileNetV3 architecture [7]. First, we compress the channel. Given the features of the proposed regenerative network, we adopt the parametric ReLU activation. To use comparative features with the previous network, we replace the average pooling layer of MobileNetV3 with a maximum pooling layer. Finally, we propose a lightweight remaking learning network, which efficiently leverages background, depth, and saliency maps for RGB-D SOD. Instead of enhancing initial saliency maps, we use these maps as input and combine them with prior information from the GDAM to remake learning. The final saliency map is obtained from the weighted sum of the previous results.

As shown in Figure 1, we use the dichotomous cross-entropy to supervise the hierarchical cross-modal principal network and the lightweight network. First, in the output of the hierarchical integration principal network, we use the reversed ground truth for supervision. Specifically, we use deep multilayer supervision over the entire lightweight network decoding branch, which maintains the feature extraction from the previous background outcome of the principal network. Second, the output of the lightweight network is processed as a multiple branching channel stack, and second-level supervision is performed to filter redundant information and reduce the excessive corrections in the lightweight network by using a simple door structure. Finally, we implement the lightweight process in the network output with the predicted results and the principal network output with background information obtained from the inversion. By weighted sum, we achieve remaking learning to generate the final saliency map. The total loss for learning can be written as

$$l_{fg} = l_{s1} + l_{s2} + l_{s3} + l_{s4} + l_{s5} + l_{s6}, \quad (1)$$

* Corresponding author (email: wujiezhou@163.com)

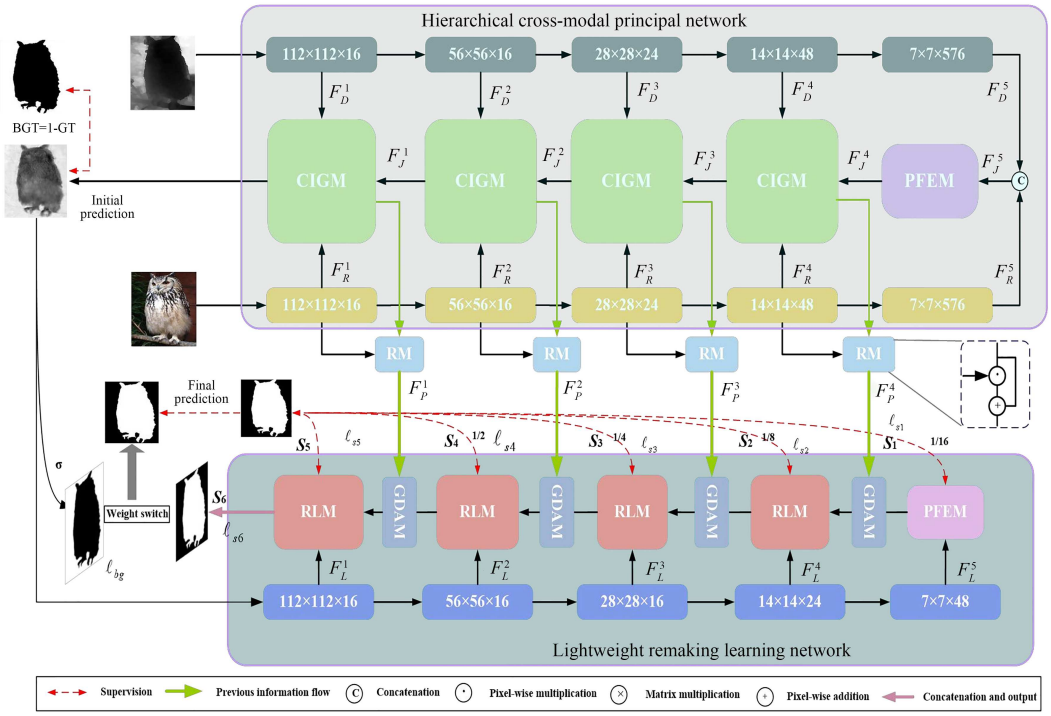

Figure 1 (Color online) Overall structure of RLLNet.

$$l_{\text{final}} = \lambda_1 l_{bg} + \lambda_2 l_{fg}, \tag{2}$$

where $l_{s1}, l_{s2}, \ldots, l_{s6}$ regularize the output of the remaking learning branch and l_{final} represents the final saliency map.

Experiments and results. We compared RLLNet with 14 state-of-the-art methods [8,9]. Because of space limitations, the detailed results are presented in Appendix C.

Conclusion. We developed a lightweight learning framework for efficient RGB-D SOD. First, a hierarchical cross-modal principal network extracts complementary depth features by learning multiscale refinement features. RGB and depth features are alternately input into the CIGM for gradual refinement. A cross-modal interaction strategy can prevent mutual degradation between RGB and depth features and achieve preliminary background noise interference. Then, the GDAM receives prior cues and imports them into a lightweight network that initializes weights. Finally, the saliency map is predicted by the weighted addition of the initial saliency map of the reversed background. Its performance on eight benchmark datasets demonstrates the effectiveness and superiority of the proposed RLLNet in terms of efficiency and robustness.

Acknowledgements This work was supported by National Natural Science Foundation of China (Grant Nos. 61502429, 61972357).

Supporting information Appendixes A–C. The supporting information is available online at info.scichina.com and link.springer.com. The supporting materials are published as submitted, without typesetting or editing. The responsibility for scientific accuracy and content remains entirely with the authors.

References

1 Zhou W J, Guo Q L, Lei J S, et al. IRFR-Net: interactive recursive feature-reshaping network for detecting salient objects in RGB-D images. IEEE Trans Neural Netw Learn Syst, 2021. doi: 10.1109/TNNLS.2021.3105484
2 Zhou T, Fan D P, Cheng M M, et al. RGB-D salient object detection: a survey. Comp Visual Media, 2021, 7: 37–69
3 Zhou W J, Wu J W, Lei J S, et al. Salient object detection in stereoscopic 3D images using a deep convolutional residual autoencoder. IEEE Trans Multimedia, 2021, 23: 3388–3399
4 Zhang Z, Lin Z, Xu J, et al. Bilateral attention network for RGB-D salient object detection. IEEE Trans Image Process, 2021, 30: 1949–1961
5 Borji A, Cheng M M, Jiang H, et al. Salient object detection: a benchmark. IEEE Trans Image Process, 2015, 24: 5706–5722
6 Zhou W J, Lv Y, Lei J S, et al. Global and local-contrast guides content-aware fusion for RGB-D saliency prediction. IEEE Trans Syst Man Cybern Syst, 2021, 51: 3641–3649
7 Howard A, Sandler M, Chu G, et al. Searching for MobileNetV3. In: Proceedings of the IEEE/CVF International Conference on Computer Vision (CVPR), 2019. 1314–1324
8 Zhang Y F, Zheng J B, Jia W J, et al. Deep RGB-D saliency detection without depth. IEEE Trans Multimedia, 2022, 24: 755–767
9 Fan D P, Lin Z, Zhang Z, et al. Rethinking RGB-D salient object detection: models, data sets, and large-scale benchmarks. IEEE Trans Neural Netw Learn Syst, 2021, 32: 2075–2089