LARNet:Towards Lightweight, Accurate and Real-time Salient Object Detection

Wang, Z., Zhang, Y., Liu, Y., Qin, C., Coleman, S. A., & Kerr, D. (2024). LARNet:Towards Lightweight, Accurate and Real-time Salient Object Detection. *IEEE Transactions on Multimedia*, *26*, 5207-5222. https://doi.org/10.1109/tmm.2023.3330082

Link to publication record in Ulster University Research Portal

Published in:
IEEE Transactions on Multimedia

Publication Status:
Published (in print/issue): 21/03/2024

DOI:
10.1109/tmm.2023.3330082

Document Version
Author Accepted version

General rights
Copyright for the publications made accessible via Ulster University's Research Portal is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The Research Portal is Ulster University's institutional repository that provides access to Ulster's research outputs. Every effort has been made to ensure that content in the Research Portal does not infringe any person's rights, or applicable UK laws. If you discover content in the Research Portal that you believe breaches copyright or violates any law, please contact pure-support@ulster.ac.uk.

Download date: 14/07/2024

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3330082

LARNet:Towards Lightweight, Accurate and Real-time Salient Object Detection

Zhenyu Wang, Yunzhou Zhang*, *Member, IEEE,* Yan Liu, Cao Qin, Sonya A. Coleman, *Member, IEEE,* and Dermot Kerr

***Abstract*—Salient object detection (SOD) has rapidly developed in recent years, and detection performance has greatly improved. However, the price of these improvements is increasingly complex networks that require more computing resources and sacrifice real-time performance. This makes it difficult to deploy these approaches on devices with limited computing resources (such as mobile phones, embedded platforms, etc.). Considering recently developed lightweight SOD models, their detection and real-time performance are always compromised in demanding practical application scenarios. To solve these problems, we propose a novel lightweight SOD method called LARNet and its corresponding extremely lightweight method LARNet* according to application requirements. These methods balance the relationship between lightweight requirements, detection accuracy and real-time performance. First, we propose a saliency backbone network tailored for SOD, which removes the need for pre-training with ImageNet and effectively reduces feature redundancy. Subsequently, we propose a novel context gating module (CGM), which simulates the physiological mechanism of human brain neurons and visual information processing, and realizes the deep fusion of multi-level features at the global level. Finally, the saliency map is output after fusion of multi-level features. Extensive experiments on popular benchmark datasets demonstrate that the proposed LARNet (LARNet*) achieves 98 (113) FPS on a GPU and 3 (6) FPS on a CPU. With approximately 680K (90K) parameters, the model has significant performance advantages over (extremely) lightweight methods, even surpassing some heavyweight models.**

***Index Terms*—lightweight, Salient object detection, Saliency backbone network, Context gating module, Feature fusion.**

I. Introduction

INSPIRED by the fact that humans can automatically and efficiently analyze complex visual scenes, computer vision algorithms should be able to quickly locate salient content and ignore other non-salient content [1]. The computer vision approach to this is salient object detection (SOD) [2]. Specifically, SOD aims to efficiently extract the important information and accurately filter out the redundant information in the visual scene, explore and simulate the visual attention mechanism of humans, assist other computer vision tasks to further extract the higher-level semantic information in the scene and to establish the understanding of the visual scene from a local to global level. In recent years, salient object detection has been widely used in applications such as object detection [3], semantic segmentation [4], RGB-D/T processing [5]–[9], simultaneous localization and mapping [10], video processing [11]–[13], robot navigation [14], person re-identification [15] and other fields [16]–[19] due to its ability to greatly reduce the complexity of subsequent processing and improving overall performance. Therefore, salient object detection has attracted much attention and flourished in the field of computer vision and image processing.

With the emergence of convolutional neural networks (CNNs) [20] and fully convolutional networks (FCNs) [21], a large number of methods based on deep learning have emerged that can achieve end-to-end salient object detection with generalization ability and detection performance far superior to traditional handcrafted methods. Recent literature [22] proposed an iterative top-down and bottom-up network for SOD, and demonstrates that most other saliency models based on FCNs are essentially variants of this model. However, the improvement in detection performance with deep learning methods means the model design is becoming increasingly more complex. In other words, models have grown in size, and the performance requirements of hardware devices has increased. For example, when the input image is 320×320, the resulting model size of MINet [23] is 650MB, the number of parameters is 162.38M, FLOPs is 87.10G, and it only runs at 41 FPS (frames per second) on a high-performance NVIDIA RTX3090 GPU. Even the speed of EGNet [24] is only 21 FPS. Obviously, such a heavyweight model requires large storage and high computing power but can only obtain poor real-time performance. These heavyweight SOD models are even difficult for high-performance devices to meet the requirements of applications in scenarios such as autonomous driving, augmented reality and real-time monitoring, and are also unsuitable for deployment on mobile devices (such as mobile phones and embedded systems).

To solve the aforementioned problems, it is important to design a lightweight SOD model that simultaneously meets the requirement to maintain detection accuracy and increase FPS. There are two major challenges for this lightweight SOD model: 1) The design of the lightweight backbone network; 2)

* Corresponding author.

Z. Wang is with the Faculty of Robot Science and Engineering, Northeastern University, Shenyang 110819, China, and also with the Technical University of Munich, 80333 Munich, Germany.(e-mail: 1910652@stu.neu.edu.cn).

Y. Zhang and C. Qin are with College of Information Science and Engineering, Northeastern University, Shenyang, China(e-mail: zhangyunzhou@mail.neu.edu.cn; qincao1994@gmail.com).

Y. Liu is with Faculty of Robot Science and Engineering, Northeastern University, Shenyang, China(e-mail: 1810630@stu.neu.edu.cn).

S. A. Coleman and D. Kerr are with the Intelligent Systems Research Centre, Ulster University, Magee Campus, Londonderry BT48 7JL, U.K.(email: sa.coleman@ulster.ac.uk; d.kerr@ulster.ac.uk).

0000–0000/00$00.00 © 2021 IEEE

Authorized licensed use limited to: UNIVERSITY OF ULSTER. Downloaded on January 19,2024 at 09:14:28 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3330082

The problem of multi-level feature fusion. With respect to the design of the lightweight network, when existing lightweight backbone networks (MobileNet-V3 [25], ShuffleNet-V2 [26], GhostNet [27], etc.) are directly applied to the field of SOD, the extracted features will be redundant, difficult to perfectly integrate with SOD tasks, and make it difficult to compress the model later. Therefore, it is necessary to build a lightweight network tailored for SOD, which not only improves the overall performance, but also eliminates the limitations of the backbone network for pre-training on ImageNet. With respect to the problem of multi-level feature fusion, how to deeply fuse the low-level information with appropriate detail needs to be considered. The integration of high-level features with accurate positioning information from the backbone network is the most important step towards produce saliency maps.

To meet the major challenges mentioned above, there are still some problems in the existing lightweight models [28]–[33]. For example, although the detection performance of iNAS-SOD [29] and DNTDF [30] has been greatly improved, their lightweight has been greatly reduced. The CSNet [33] has a high degree of lightweight, but its detection performance is weak. They difficult to balance the relationship between detection performance and computing resources. Therefore, we propose a lightweight, accurate and real-time network for SOD, named LARNet. Meanwhile, we also propose LARNet* for an extremely lightweight SOD model according to different application requirements. As shown in Figure 1, the details are as follows:

Firstly, considering the problem of a model's lightweight nature and feature redundancy, different from Li et al's [19] lightweight VGG-16 for building subnetworks. Jin et al. [8] designed an asymmetric dual-stream encoders based on MobileNet V3. Huang et al. [34] proposed an LD-ResNet-18 backbone based on ResNet-18, while Gu et al. [29] achieved the best performance-latency balance with the help of an integral neural architecture search. Liu et al. were inspired by cognitive science to design a backbone consisting of HVP [32] or SAM [31], and Cheng et al. [33] designed a generalized OctConv to build a backbone. We propose a new approach: the saliency backbone network replaces the complex construction method with a direct construction method. We only build the saliency backbone network through (depth-wise separable) convolution without adding other enhancement modules. Therefore, we propose two backbone networks: a lightweight saliency backbone network LSBNet and an extremely lightweight saliency backbone network ELSBNet tailored for SOD tasks, which can improve the overall network performance without pre-training on ImageNet. Meanwhile, as a relatively independent backbone network, either of these can replace the backbone network in existing SOD methods, which demonstrates strong flexibility.

Next improvement focuses on fusing the multi-level features of the backbone network output. Different from heavyweight SOD methods, the multi-level feature fusion of lightweight SOD methods needs an efficient fusion mechanism, which can achieve accurate performance with less network parameters. In cognitive science, there is a large number of "excitatory neurons" and "inhibitory neurons" in the human brain. Presynaptic neurons that increase the firing rate of postsynaptic neurons are "excitatory neurons", and the "inhibitory neurons" decrease the firing rate. The interaction between excitatory and inhibitory neurons allows humans to quickly obtain important information [35]. In the process of visual perception, humans initially have overall cognition of the global environment, and then can switch their attention to a salient object [36]. Inspired by this, we believe that in the field of SOD, this attention mechanism can be well simulated by the gating module, which is equivalent to setting up an information transmission mechanism to coordinate the interaction between "excitatory neurons" and "inhibitory neurons". Meanwhile, the gated module is endowed with global perception capabilities. Therefore, we propose a lightweight context gating module (CGM) to achieve feature fusion between multi-level features at the global level. Finally, the lightweight feature fusion approach is used to decode the features of CGM effectively and output the saliency map, and then use a loss function to optimize the corresponding predicted saliency map at the pixel and object level.

In summary, the main contributions are as follows:

1) We propose a(n) (extremely) lightweight, accurate and real-time SOD method, named LARNet (LARNet*) which has a good balance between being lightweight, detection accuracy and real-time performance.
2) We propose lightweight saliency backbone networks LSBNet and ELSBNet tailored for lightweight SOD, which maintain good performance without pre-training on ImageNet, and have better portability.
3) We propose a novel context gating module (CGM), which effectively enriches the features of all levels through global information transmission, and simulates the brain-inspired excitation mechanism efficiently. We also include a lightweight feature fusion approach, which decodes multi-level features in a gradual manner.
4) The proposed LARNet (LARNet*) has reached a high level of detection performance with 0.66M (0.09M) model parameters, and using GPU and CPU achieves 98 (113) and 3 (6) FPS, respectively. Compared with other state-of-the-art methods, the output saliency map shows superiority using benchmark datasets.

The remainder of the paper is organized as follows. Section II reviews the state-of-the-art salient object detection methods, including heavyweight and lightweight methods. Section III presents our proposed LARNet (LARNet*), describes its network architecture and the important modules. Section IV verifies the superiority and effectiveness of our proposed method through comparative experiments with other state-of-the-art methods and Section V summarizes the paper.

II. Related work

Visual saliency detection can be traced back to 1998 and was proposed by Itti et al [37]. Subsequently, Lai et al. [38] conducted systematic research on the use of artificial and human attention in neural network design, and demonstrated through experiments that human attention is valuable for achieving better performance in deep networks and enhancing

Authorized licensed use limited to: UNIVERSITY OF ULSTER. Downloaded on January 19,2024 at 09:14:28 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3330082

robustness to disturbances. After more than 20 years of development, the research is mainly divided into two categories: traditional methods [39]–[49] based on handcrafted features and deep learning methods [22]–[24], [28]–[30], [30]–[33], [50]–[77] based on high-level semantic features. The traditional methods mainly rely on information such as color, texture and priori center. Although traditional approaches can achieve good prediction results, they are difficult to apply in practice due to their inability to detect a complete salient object and their poor ability to suppress noise with complex foreground or background environments. Deep learning methods based on high-level semantic features can effectively solve the above problems, and have shown explosive growth in recent years. In this paper, we focus on the deep learning based methods.

A. Heavyweight Salient Object Detection

Most current salient object detection research focuses on achieving good performance by fusing the multi-level features output from the backbone network, which is generally based on ResNet [78] or VGG [79]. Scholars have designed various networks and strategies to fuse multi-level features to obtain accurate saliency maps. For example, Wu et al. [73] proposed a multi-task algorithm for SOD, foreground contour detection, and edge detection to alleviate the problem of incomplete saliency maps. An intertwined supervision strategy is adopted, and the proposed mutual learning module effectively improves the performance of the network, achieving a more accurate saliency map while detecting satisfactory edges. Wang et al. [74] proposed an attentive saliency network that connects fixation and SOD, learning to detect salient objects from fixations, and narrowing the gap between SOD and fixation prediction. Wei et al. [61] proposed to decouple the saliency label into body mapping and detail mapping. They make full use of the complementarity of body mapping and detail mapping to generate high-quality saliency maps. Li et al. [66] proposed a stacked U-type network with channel-wise attention, which is composed, in parallel, of a dilated convolution module and a multi-level attention cascade feedback module. It can effectively avoid the gridding problem and can describe the inter-dependence between different channel maps in the same layer. Xu et al. [67] simulated human biological capabilities and proposed a progressive architecture with a knowledge review network to make full use of the information of each layer by recombining the finest feature maps with those from previous layers. Zhuge et al. [68] designed the diverse feature aggregation module, the integrity channel enhancement module, and the part-whole verification module. By integrating these modules, the proposed ICON can capture diverse features at each feature level and enhance feature channels. Yao et al. [71] focused on edge problems and proposed a saliency detection unit to learn more boundary features, and apply multiple such units to construct a boundary information progressive guidance network. Then, a boundary information guidance module is designed, which focuses on the boundary information in the feature layer. Liu et al. [76] proposed a novel disentangled part-object relational (POR) network, and also proposed a residual learning method to integrate contrast cues and POR cues for saliency prediction. In addition, SOD is also widely used in video, RGB-D and other fields. Fu et al. [7] proposed two effective components of joint learning and densely cooperative fusion, which achieved cross-modal efficient fusion of RGB image and depth, providing new insights for RGB-D SOD. Fan et al. [18] proposed a CoEG-Net that augments the EGNet model with a co-attention projection strategy for fast common information learning, enabling a study on the co-salient object detection problem for images.

Although high detection accuracy is obtained, the resulting large model is difficult to be applied to actual scenes. For example, the current state-of-the-art method PA-KRN [67] has a model size of 790.8MB, which requires considerable computing power and has low real-time performance. It is almost impossible to deploy on practical systems. Even the relatively small model LDF [61] requires 100.9MB, which is still challenging to deploy on practical systems.

B. Brain-inspired Networks

Recently, due to the increased interest in human cognitive science and artificial intelligence, many brain-inspired networks emerged in the field of artificial intelligence. For example, inspired by the mammalia brain that can effectively solve catastrophic forgetting by consolidating memory as more specific or generalized form to complement each other, Wang et al. [80] proposed a triple-memory network (TMN) for continual learning, and realized state-of-the-art performance of generative memory replay. To further improve the performance of a multilayer perceptron (MLP), Li et al. [81] combined a brain-inspired spiking neural network (SNN) with a MLP, enabling the overall network to achieve higher accuracy without extra FLOPs. Inspired by the knowledge of neuroscience, Chang et al. [82] developed a memory formation system (MFS) to establish memory for a GAN, simulating human encoding, consolidation, and retrieval functions in memory formation, effectively addressing catastrophic forgetting problems. Inspired by the dynamic plasticity of dendritic spines, Zhao et al. [83] proposed a brain-inspired developmental neural network based on dendritic spine dynamics (BDNN-dsd) which simulates their behaviours and can improve the network convergence speed and classification performance even for compact networks. Li et al. [84] proposed a hybrid loop closure detection (LCD) method based on convolutional neural network features and a locality-sensitive hashing algorithm to solve the problem of challenging or large-scale environments that existing brain-inspired SLAM system LCD methods cannot effectively solve through manually crafted features and brute force search strategy. This enables the system to construct cognitive maps with better robustness and efficiency. For the SOD field, inspired by the primate visual system's hierarchical processing of visual signals with different receptive fields and eccentricities in different visual cortex areas, Liu et al. [32] proposed a hierarchical visual perception (HVP) module to imitate the primate visual cortex for hierarchical perception learning, and improved the overall performance of the model.

The recently popular attention mechanism is also an important part of brain-inspired research, which achieves fo-

Authorized licensed use limited to: UNIVERSITY OF ULSTER. Downloaded on January 19,2024 at 09:14:28 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3330082

cused attention on key objects in the perceptual environment. Specifically, Woo et al. [85] exploited the inter-channel/spatial relationships and adaptively recalibrated the feature map in a channel/spatial manner, emphasizing the features of important objects in the perceptual environment. For the SOD field, Liu et al. [31] proposed a stereoscopic attention mechanism to adaptively recalibrate the feature flow from multiple branches by means of channel and spatial clues, and realized the lightweight nature of the SOD model. It is worth noting that Lai et al. [75] proposed a weakly supervised method for visual saliency prediction. They modeled a set of cognitive theories of visual attention as network modules, including spatial visual semantics, object-related cues, winner-take-all theory and center priors, achieving high performance.

In summary, it is clear that the introduction of brain-inspired networks can effectively improve the comprehensive performance of various tasks and make the network interpretable to a certain extent.

C. Lightweight Salient Object Detection

With the continuous improvement in network detection performance, the size of the resulting model is significantly increasing, and thus the real-time performance is seriously affected. These models need more storage space and higher computing power, which is contrary to the requirements of real-world applications. Therefore, lightweight SOD models have received more attention in recent years. Gao et al. [33] constructed an extremely lightweight model, CSNet, and proposed a generalized OctConv (gOctConv). Combined with a dynamic weight decay scheme, the saliency map can be achieved with only approximately 100k model parameters and it can be trained directly from scratch without ImageNet pre-training. Liu et al. [32] proposed a hierarchical visual perception (HVP) module and built a lightweight backbone network for SOD with the help of Conv, DSConv, HVP module, attention and a dropout mechanism, but it requires pre-training with ImageNet to achieve the best results. Compared with CSNet, the detection performance of HVPNet has greatly improved using 1.24M model parameters. Subsequently, Liu et al. [31] again proposed a novel stereoscopically attentive multiscale (SAM) module that enables small networks to efficiently encode both high-level features and low-level details. It uses Conv, DSConv, SAM and PPM modules to build a lightweight backbone network for SOD, which also requires pre-training on ImageNet to achieve optimal results. The number of model parameters for SAMNet is 1.33M, and its detection performance is similar to HVPNet. Recently, Wu et al. [29] proposed a device-aware search scheme, which trains the SOD model only once and achieves high-performance but low-latency on multiple devices. With only 4.96M model parameters, this scheme achieves the best detection performance in the field of lightweight models. Subsequently, Fang et al. [30] designed a novel framework based on densely nested top-down flows (DNTDF). Integrating DNTDF with EfficientNet, a SOD model with only 4.61M model parameters was built, and it showed strong detection performance. Finally, Wu et al. [28] proposed an Extremely-Downsampled Network (EDN), which uses extreme sub-sampling technology to effectively learn the global view of the whole image. Among them, EDN-Lite has reached a high detection performance with 1.8M parameters. When we compare the inference speed on an NVIDIA RTX3090 GPU and crop the input image to 320×320, the inference speed of CSNet [33], HVPNet [32], SAMNet [31], DNTDF [30] and EDN-Lite [28] are 48FPS, 43FPS, 31FPS, 61FPS, 55FPS respectively, which are all slower than the heavyweight model LDF (69FPS). In addition, for RGB-thermal SOD, Zhou et al. [9] proposed a lightweight spatial boosting network, in which the boundary boosting algorithm can optimize the predicted saliency map and reduce the information collapse in low-dimensional features, which relies on 5.39M parameters and achieves competitive performance.

As mentioned above, although the lightweight SOD model has achieved some good results, it still requires further development. Different from other methods, we propose a Context Gating Module, combined with our lightweight saliency backbone network. The overall approach achieves the best results in terms of being lightweight, accurate and running in real-time, and promotes the progress of its deployment and application.

III. Proposed Method

A. The Overall Architecture

We propose a novel salient object detection network LARNet, which is aimed at being lightweight, accurate and run in real-time. Simultaneously, we propose LARNet* as an extremely lightweight model. The overall architecture of the network is shown in Figure 1. It is worth noting that we use Conv to represent the conventional convolution operation and DSConv [25] to represent the depthwise separable convolution operation. Batch normalization and rectified linear unit are performed once after each convolution. Like other lightweight models [28], [31]–[33], in order to reduce the computational requirements of the model as much as possible, DSConv is used to perform convolution. The only input is an RGB image, and it is input into the lightweight saliency backbone network (LSBNet) or the extremely lightweight saliency backbone network (ELSBNet) to obtain multi-level features with a uniform number of channels (see Section III.B). For the convenience of description, as shown in Figure 1, blocks of different colors output information streams of different colors. Multi-level feature information streams are input to our proposed context gating module (CGM) (see Section III.C), which outputs useful information after multi-level feature fusion. Then we perform feature fusion and decode between adjacent features in a step-by-step manner, and output saliency maps (see Section III.D). Finally, We use the binary cross entropy (BCE) + intersection-over-union (IOU) hybrid loss function to fully supervise the output saliency map at each level of the network, so that the limited parameters can be learned to optimise the information. The experimental results in Section IV also demonstrate that our method has greatly improved performance compared with other state-of-the-art approaches.

B. Lightweight Backbone Network

To obtain a lightweight model, popular lightweight backbone networks (such as MobileNet-V3 [25], ShufflNet-V2

Authorized licensed use limited to: UNIVERSITY OF ULSTER. Downloaded on January 19,2024 at 09:14:28 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3330082

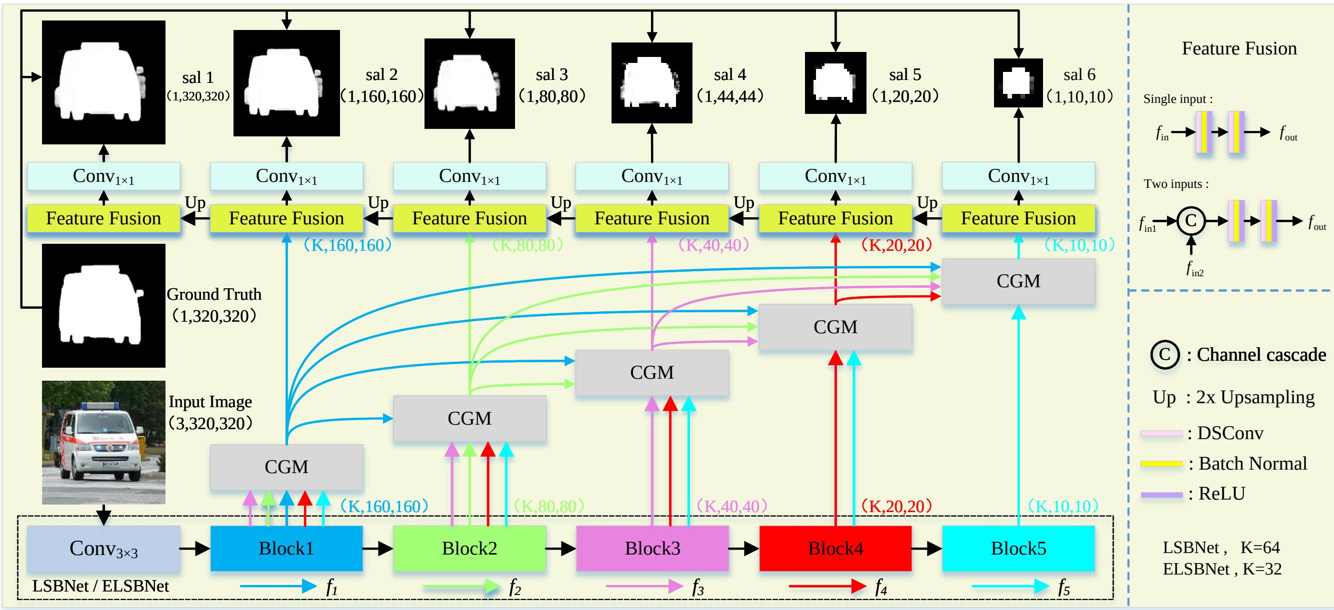


Fig. 1. The overall architecture of LARNet and LARNet*. The difference between them lies in the choice of saliency backbone network. The former is LSBNet, while the latter is ELSBNet. There is only a difference in the number of feature channels between them.

[26], and GhostNet [27]) have been introduced to replace heavyweight backbone networks (such as ResNet [78], VGG [79]) in SOD models. However, there are still two disadvantages: 1) it is difficult to deploy these lightweight networks on equipment with limited resources; 2) they are ependent on an ImageNet pre-trained model, resulting in feature redundancy and hindering further compression of the model.From Section II.B, we can see that CSNet [33], HVPNet [32] and SAMNet [31] have designed complex backbone networks. These backbone networks are excellent, however we believe that a lightweight backbone network should be simple and reduce the insertion of additional modules. Therefore we present a new idea, that is, only Conv and DSConv are used to build an efficient backbone network and high performance can be achieved without pre-training with ImageNet to overcome the two disadvantages.

TABLE I
SALIENCY BACKBONE NETWORK SETTINGS OF THE PROPOSED LSBNET AND ELSBNET. N REPRESENTS THE NUMBER OF MODULES. OC REPRESENTS THE NUMBER OF OUTPUT CHANNELS OF THE MODULE. S REPRESENTS STRIDE. I REPRESENTS THE MULTIPLICATION FACTOR OF THE INPUT CHANNEL. WHERE THE PARAMETERS FROM ELSBNET DIFFER FROM LSBNET THESE ARE DENOTED IN BRACKETS.

Stage	Input	Module	N	OC	S	I
1	$320^2 \times 3$	$\text{Conv}_{3\times3}$	1	32	2	-
	$160^2 \times 32$	Block	1	64(32)	1	1
2	$160^2 \times 64(32)$	Block	1	64(32)	2	6(2)
	$80^2 \times 64(32)$	Block	1	64(32)	1	6(2)
3	$80^2 \times 64(32)$	Block	1	64(32)	2	6(2)
	$40^2 \times 64(32)$	Block	1	64(32)	1	6(2)
4	$40^2 \times 64(32)$	Block	1	64(32)	2	6(2)
	$20^2 \times 64(32)$	Block	2	64(32)	1	6(2)
5	$20^2 \times 64(32)$	Block	1	64(32)	2	6(2)
	$10^2 \times 64(32)$	Block	2	64(32)	1	6(2)

To meet different requirements, we propose a lightweight saliency backbone network (LSBNet) and extremely lightweight saliency backbone network (ELSBNet). The specific settings of LSBNet and ELSBNet are shown in Table I. The difference between the two is only in the configuration of "OC" and "I", and ELSBNet is a lighter version of LSBNet. They output 5-level features like other backbone networks. The formula of each block is defined as:

$$Block = \begin{cases} x + Conv_{1\times1}(DwConv_{3\times3}(Conv_{1\times1}(x))), & stride = 1 \\ Conv_{1\times1}(DwConv_{3\times3}(Conv_{1\times1}(x))), & stride = 2 \end{cases} \tag{1}$$

where x is the input feature.

It can be seen that in the saliency backbone network we propose, each component plays a key role. As shown in Table 1, the $\text{Conv}_{3\times3}$, and each block in the module column, form the backbone network in the form of a cascade. Firstly, we use $\text{Conv}_{3\times3}$ to extract features from the input image, which effectively reduces the information loss from the original image. Then, the whole structure uses a block with a stride of 2 to achieve feature down-sampling, and after each down-sampling operation, a block with a stride of 1 is used to achieve further feature extraction and enhancement. For stages 1 to 3, the feature resolution is relatively high, so only one block with a stride of 1 is used to extract and enhance the features after down-sampling, which effectively reduces the computational complexity. For stages 4 to 5, the feature resolution is relatively low, so two blocks with a stride of 1 are cascaded after down-sampling, which can enhance the richness of high-level semantic information while not significantly increasing the amount of computation. It is worth noting that the novelty of our saliency backbone network is that, except for the first convolution layer which outputs 32 channels, all other blocks output 64(32) channels. This has two main advantages:

Authorized licensed use limited to: UNIVERSITY OF ULSTER. Downloaded on January 19,2024 at 09:14:28 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3330082

1) The post-processing does not need to unify the number of channels for the output of the saliency backbone network.

2) For lightweight backbone networks, the number of output channels for low-level features is often less than 64(32). Here, the low-level features can be extracted more abundantly. Although some information will be lost for the extraction of high-level features, it can achieve a good balance between lightweight and detection performance for SOD tasks.

The selection of the "I" value in Table I is obtained through experiments. Generally a larger "I" value can make the network learn more features, but it is accompanied by an increase in model complexity and even feature redundancy. In the ablation experiment in Section IV, we also carried out relevant verification and confirmed our observations. As far as we know, LSBNet (ELSBNet) is the simplest and most efficient saliency backbone network in the (extremely) lightweight SOD field.

C. Context Gating Module

We consider multi-level feature fusion to be particularly important. We hope that in the process of multi-level feature fusion, the network can recover and learn more useful information. As described in the introduction, inspired by the physiological mechanisms of human brain neurons [35] and visual information processing [36], we propose a novel context gating module (CGM). For ease of understanding, we use the CGM corresponding to Block1 (as shown in Figure 1) as an example, as illustrated in Figure 2. This is a lightweight and efficient module, which realizes the deep fusion between features at a global level. The working mechanism of CGM mainly includes three stages, as follows:

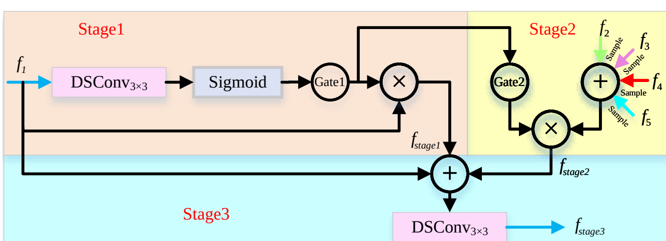


Fig. 2. An overview of proposed context gating module (CGM). It consists of three stages, simulating and realizing a mechanism inspired by the brain.

Stage 1: the input feature f_1 is the output feature of the block (as shown in Figure 1) in the same layer as the CGM. The input features f_1 are defined as "main features" of the module, and DSConv is used to learn it. Then, the feature values are normalized to $[0, 1]$ by a Sigmoid function, and "Gate1" is formed. The purpose of "Gate1" is to highlight the useful information in the "main features" f_1 and simulate the "excitatory neurons" of the human brain. "Gate1" and the final output f_{stage1} of the first stage are expressed as:

$$f_{stage1} = Gate1 \times f_1, Gate1 = Sigmoid(DSConv_{3\times3}(f_1)) \quad (2)$$

Stage 2: the input features f_2, f_3, f_4, f_5 are the output features of the blocks (as shown in Figure 1) in different layers from the CGM. To reduce the computational complexity, we sample the features f_2 to f_5 respectively to be the same size as the feature f_1 and add them directly, and the fused input is defined as "secondary features", giving the feature global observability. Through the "Gate2" mechanism, the "secondary features" can be used to supplement the foreground features when the "main features" f_1 have not successfully extracted (for example, the information loss caused by the sampling process in the backbone network). At the same time, the background features of the "main features" f_1 are obtained from the "secondary features", effectively simulating the "inhibitory neurons" of the human brain. "Gate2" and the final output f_{stage2} of the second stage are expressed as:

$$f_{stage2} = Gate2 \times (\sum_{i=2}^{5} Sample(f_i)), \quad Gate2 = 1 - Gate1 \quad (3)$$

where $Sample()$ uses the $interpolation(``bilinear")$ function in the Pytorch library.

Stage 3: f_{stage1}, f_{stage2} and f_1 are combined in an additive manner to reduce the computational complexity. The features f_{stage1} as the brain-inspired "excitatory neurons" and the features f_{stage2} as the brain-inspired "inhibitory neurons" interact with each other to strengthen the "main features" f_1. Finally, the features are further learned by DSConv and the fused features are output, realizing the tight coupling of features at the global level. The final output f_{stage3} of the third stage is expressed as:

$$f_{stage3} = DSConv_{3\times3}(\sum_{i=1}^{2} f_{stage(i)} + f_1) \quad (4)$$

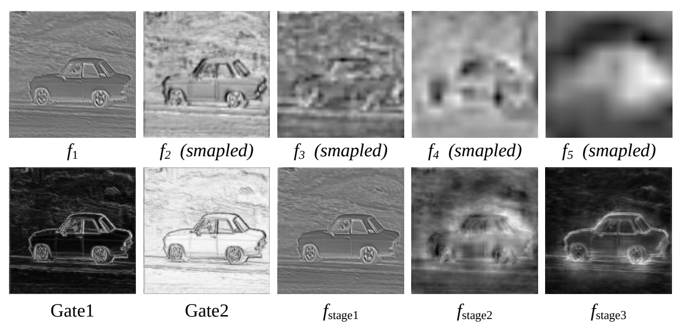


Fig. 3. Visualization of CGM. The visualized feature maps have been added in the channel dimension (the number of channels in the visualized feature maps is 1). The first row is the input features of CGM in stage 1 and 2. The second row is the output features of Gate1, Gate2 and each stage in CGM.

As illustrated in the CGM visualization in Figure 3 (The symbols of the features displayed in Figure 3 correspond to Figure 2.), it is clear to see that simulation of "excitatory neurons" and "inhibitory neurons" is achieved using Gate1 and Gate2, achieving interaction and fusion of "secondary features" and "main features" similar to the human brain. Comparing f_1 and f_{stage3}, CGM highlights the salient object (car) in the image and pays attention to the edges of the salient object (car). The context gating module (CGM) we proposed has two main advantages:

1) The connection between multi-level features is effectively utilized, and the ingenious fusion of local and global features is achieved.

2) It is a plug-and-play module, which achieves high performance with a simple and lightweight architecture.

In the ablation experiment in Section IV, we also demonstrate the superiority and necessity of this module.

Authorized licensed use limited to: UNIVERSITY OF ULSTER. Downloaded on January 19,2024 at 09:14:28 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3330082

TABLE II
THE DATASETS AND EVALUATION CRITERIA FOR SALIENT OBJECT DETECTION

Datasets					
Name	Year	Stage	Size	Characteristic	Attribute
DUTS-TR [86]	2017	Train	10553	Complex	Multi-object, different sizes
DUTS-TE [86]	2017	Test	5019	Complex	Multi-object, different sizes
DUT-OMRON [87]	2013	Test	5168	Complex	Multi-object, different sizes
ECSSD [88]	2015	Test	1000	Simple	Mostly single-object, large size
PASCAL-S [89]	2014	Test	850	Complex	Multi-object, moderate size
HKU-IS [90]	2015	Test	4447	Complex	Multi-object, moderate size

Evaluation Criteria		
Name	Formula	Characterization
F-measure [91]	$mF = \frac{(1+\beta^2)Precision*Recall}{\beta^2 Precision+Recall}$	Weighted combination of precision and recall
Mean Absolute Error [92]	$MAE = \frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\lvert P(i,j)-G(i,j)\rvert$	Average absolute difference between the output and the GT
Enhanced-alignment measure (E_ξ) [93]	$E_\xi = \frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\phi s(i,j)$	Global means of the image and local pixel matching simultaneously
Structural measure (S_α) [94]	$S_\alpha = \alpha \times S_o + (1-\alpha)\times S_r$	Similarity Evaluation of the regional and object perception structure
Intersection-over-Union	$IOU = \frac{TP}{TP+FP+FN}$	Overlap between the output and the GT
Model Parameters	#Param	Lightweight degree of the model
Model Size	#Size	Size of storage occupied by the model
Floating-Point Operations	FLOPs	Computational cost of the model
Frames Per Second	FPS	Real-time performance of the model

D. Feature Fusion

As shown in Figure 1, the resolution of the features output at each level of the CGM is different. As a lightweight SOD model, we still adhere to the use of a simple and efficient method to solve this problem. Therefore, as shown in Figure 1, we only use DSConv to decode and fuse features as follows:

1) When there is only a single input, the approach mainly realizes a more comprehensive extraction of input features, effectively avoiding the loss of useful information caused by the subsequent up-sampling operation. The output f_{out} is expressed as:

$$f_{out} = DSConv_{3\times3}(DSConv_{3\times3}(f_{in})) \tag{5}$$

2) When there are two inputs, the approach mainly realizes the deep fusion of the features from the two inputs, and also reduces the loss of useful information caused by the subsequent up-sampling operation. The output f_{out} is expressed as:

$$f_{out} = DSConv_{3\times3}(DSConv_{3\times3}(concat((f_{in1}, f_{in2}), dim = C))) \tag{6}$$

In the ablation experiment in Section IV, we also prove the superiority of this module. Through the combination of all components, our proposed method obtains the best results in the trade-off among lightweight, accuracy and real-time.

IV. EXPERIMENTS

A. Experimental Preparation

Table II lists and describes the datasets and evaluation criteria used.

B. Implementation Details

We train the model using the DUTS-TR dataset and adapt data augmentation techniques (such as horizontal flip, random crop and multi-scale input images) to increase the training dateset. Similar to ICON [68], we use binary cross entropy (BCE) loss and intersection over union (IOU) loss to jointly supervise the network. Our training loss L_{toal} is defined as $L_{toal} = \sum_{i=sal1}^{sal6} L_i, L_i = L_{bce} + L_{iou}$. Our model is built on the PyTorch platform and runs on an NVIDIA RTX3090 GPU or Intel(R) Xeon(R) Platinum 8157 CPU @ 2.30GHz. The network is trained end-to-end by stochastic gradient descent (SGD), and the momentum and weight decay are set to 0.9 and 0.0005, respectively.

The training of the network is divided into two steps: at the first step, we train our initial model. At the second step, we map the initial trained model parameters to the final model for training. The training process is completed in one shot, and the initial parameter mapping process is completed automatically during the period. The difference between the initial model and the final model is only the number of modules, parameter $N = \{1, 1, 1, 1, 1, 2, 1, 6, 1, 3\}$ in the saliency backbone network, the remaining parameters are consistent. This training strategy can provide better initial parameters for LARNet (LARNet*), which is more conducive to speeding up convergence and improving stability, and is more conducive to obtaining the best performance. In both stages, warm-up and linear decay strategies are used [61], the maximum learning rate for each is set to 0.05 and 0.005, respectively. The batchsize is set to 32, and the maximum periods are set to 100 and 200, respectively, and the total training time of LARNet (LARNet*) is about 17 (13) hours. During testing, each image is resized to 320×320 pixels and then fed into the network to obtain a prediction, and finally restored to the original image size through bilinear interpolation [61], [68].

C. Performance comparison

We compare our model with 26 state-of-the-art SOD methods, which are BASNet [51], CPD [52], PoolNet [53], SCRN [54], EGNet [24], DFI [55], U2-Net [56], GCPANet [57], F3Net [58], GateNet [59], ITSD [60], MINet [23], LDF [61], PSGL-Net [95], Auto-MSFNet [96], VST [97], PFSNet [98], ICON [68], OLER [69], HVPNet [32], SAMNet [31], iNAS-SOD [29], DNTDF [30], EDN-Lite [28], CSNet [33] and CSNet* [33]. The comparison of heavyweight methods only emphasizes that they require higher computing resources. As this paper is aimed at (extremely) lightweight methods, it mainly aims at a comprehensive comparison of the

Authorized licensed use limited to: UNIVERSITY OF ULSTER. Downloaded on January 19,2024 at 09:14:28 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3330082

TABLE III

DETECTION PERFORMANCE COMPARISON WITH 26 STATE-OF-THE-ART METHODS USING FIVE DATASETS. MF (LARGER IS BETTER), MAE (SMALLER IS BETTER). THE BEST RESULTS OF LIGHTWEIGHT METHOD AND EXTREMELY LIGHTWEIGHT METHOD ARE MARKED WITH BOLD **RED**. TO ENSURE FAIRNESS, WE UNIFORMLY CROP THE INPUT IMAGE TO 320×320 RESOLUTION (EXCEPT FOR VST WHERE THE IMAGE SIZE IS 224×224 AND AUTO-MSFNET WHERE THE IMAGE SIZE IS 256×256, AND DNTDF MEASURES FLOPS WITH IMAGE SIZE OF 288×288.) AND RUN IT ON THE SAME GPU AND CPU. IN THE FPS COLUMN, THE NUMBERS IN PARENTHESES ARE THE RESULTS OBTAINED USING THE CPU.

Methods	Year	#Param (M)	#Size (MB)	FLOPs (G)	FPS	DUTS-TE 5019 images		DUT-OMRON 5168 images		ECSSD 1000 images		PASCAL-S 850 images		HKU-IS 4447 images	
						mF↑	MAE↓	mF↑	MAE↓	mF↑	MAE↓	mF↑	MAE↓	mF↑	MAE↓
Heavyweight method (#Param>10M)															
BASNet [51]	CVPR 2019	87.06	348.5	199.31	32	.791	.048	.756	.056	.880	.037	.771	.076	.895	.032
CPD [52]	CVPR 2019	42	192.0	14.73	46	.805	.043	.747	.056	.917	.037	.820	.071	.891	.034
PoolNet [53]	CVPR 2019	69.56	278.5	89.65	45	.819	.037	.752	.054	.919	.035	.826	.065	.903	.031
SCRN [54]	ICCV 2019	25.23	101.4	12.53	38	.809	.040	.746	.056	.918	.037	.827	.063	.896	.034
EGNet [24]	ICCV 2019	111.66	447.1	244.13	21	.815	.039	.755	.053	.920	.037	.817	.074	.902	.031
DFI [55]	IEEE TIP 2020	29.61	118.8	22.44	42	.814	.039	.752	.055	.920	.035	.830	.065	.902	.031
U2-Net [56]	PR 2020	44.01	176.3	58.83	45	.792	.045	.761	.054	.892	.033	.770	.074	.896	.031
GCPANet [57]	AAAI 2020	67.06	268.6	54.36	58	.817	.038	.748	.056	.919	.035	.827	.062	.898	.031
F3Net [58]	AAAI 2020	25.54	102.5	13.63	65	.840	.035	.766	.053	.925	.033	.835	.061	.910	.028
GateNet [59]	ECCV 2020	128.63	514.9	112.64	36	.807	.040	.746	.055	.916	.040	.819	.067	.899	.033
ITSD [60]	CVPR 2020	26.07	106.2	19.71	53	.804	.041	.756	.061	.895	.034	.785	.066	.899	.031
MINet [23]	CVPR 2020	162.38	650.0	87.10	41	.828	.037	.755	.056	.924	.033	.829	.064	.909	.029
LDF [61]	CVPR 2020	25.15	100.9	12.87	69	.855	.034	.773	.052	.930	.034	.843	.060	.914	.028
PSGL-Net [95]	IEEE TIP 2021	25.55	102.6	16.12	61	.849	.036	.772	.053	.932	.031	.842	.061	.917	.028
Auto-MSFNet [96]	ACM MM 2021	33.35	130.4	24.55	58	.856	.034	.778	.050	.929	.033	.843	.061	.914	.027
VST [97]	ICCV 2021	44.09	178.4	23.24	38	.818	.037	.756	.058	.920	.033	.829	.061	.900	.029
PFSNet [98]	AAAI 2021	31.18	125.1	37.61	40	.846	.036	.774	.055	.932	.031	.837	.063	.919	.026
ICON [68]	IEEE TPAMI 2022	33.04	132.8	17.33	60	.838	.037	.772	.057	.928	.032	.833	.064	.910	.029
OLER [69]	ESWA 2022	26.58	106.7	-	-	.866	.033	.792	.050	.937	.030	.843	.063	.924	.026
Lightweight method (10M>= #Param>500K)															
HVPNet [32]	IEEE TCYB 2020	1.24	5.3	1.05	43 (1)	.749	.058	.721	.065	.889	.052	.784	.089	.872	.044
SAMNet [31]	IEEE TIP 2021	1.33	5.8	**0.50**	31 (1)	.745	.058	.717	.065	.891	.050	.778	.092	.871	.045
iNAS-SOD [29]	ICCV 2021	4.96	20.6	0.90	**98** (**8**)	**.809**	.039	.746	.054	**.917**	.037	**.821**	.064	**.898**	.032
DNTDF [30]	SCIS 2022	4.61	55.1	0.79	61 (6)	.806	**.035**	**.751**	**.052**	.899	**.034**	.795	**.063**	**.898**	**.030**
EDN-Lite [28]	IEEE TIP 2022	1.80	7.7	0.75	55 (7)	.781	.050	.739	.058	.897	.049	.799	.084	.883	.040
Ours	year	**0.66**	**3.0**	3.77	**98** (3)	.793	.052	.745	.065	.907	.041	.801	.082	.895	.036
Extremely Lightweight method (#Param)<=500K)															
CSNet [33]	IEEE TPAMI 2021	0.14	0.7	1.46	48 (1)	.687	.074	.675	.081	.844	.065	.723	.103	.840	.059
CSNet* [33]	IEEE TPAMI 2021	**0.09**	**0.5**	0.89	48 (2)	.666	.082	.656	.087	.831	.074	.717	.111	.826	.065
Ours*	year	**0.09**	0.6	**0.82**	**113**(6)	**.727**	**.069**	**.694**	**.080**	**.867**	**.055**	**.759**	**.096**	**.862**	**.046**

state-of-the-art lightweight methods. For fair comparison, we use the implementations with the recommended parameters and the saliency maps with the best performance provided by the authors, and the lightweight methods were tested using the same hardware. It is worth noting that due to the different evaluation implementations, the detection performance metrics in many papers show different values. To ensure fairness, we used the evaluation code provided by **https://github.com/jiwei0921/Saliency-Evaluation-Toolbox** to compare the detection performance of all methods.

1) Quantitative Comparison: As shown in Table III, according to the number of model parameters, we divide methods into three categories: heavyweight methods (#Param>10M), lightweight methods (10M>= #Param>500K) and extremely lightweight methods (#Param)<=500K). This paper mainly focuses on lightweight methods, but since extremely lightweight methods have the same status, we propose a lightweight model LARNet and an extremely lightweight model LARNet*, and compare them with other state-of-the-art methods. To prove the more powerful performance and generalization ability of our method, we evaluate using five well-known datasets, and the evaluation criteria were divided into three aspects: detection performance, efficiency performance and comprehensive performance.

TABLE IV

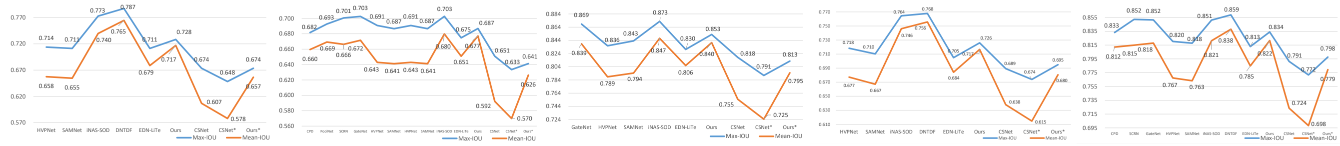
DETECTION PERFORMANCE COMPARISON WITH 26 STATE-OF-THE-ART METHODS USING FIVE DATASETS. E_ξ (LARGER IS BETTER), S_α (LARGER IS BETTER). THE BEST RESULTS OF LIGHTWEIGHT METHOD AND EXTREMELY LIGHTWEIGHT METHOD ARE MARKED WITH BOLD **RED**.

Methods	DUTS-TE 5019 images		DUT-OMRON 5168 images		ECSSD 1000 images		PASCAL-S 850 images		HKU-IS 4447 images	
	E_ξ↑	S_α↑	E_ξ↑	S_α↑	E_ξ↑	S_α↑	E_ξ↑	S_α↑	E_ξ↑	S_α↑
Heavyweight method (#Param>10M)										
BASNet [51]	.884	.866	.869	.839	.921	.916	.853	.838	.946	.909
CPD [52]	.886	.869	.866	.825	.925	.918	.855	.848	.944	.905
PoolNet [53]	.896	.887	.868	.831	.925	.926	.859	.865	.951	.918
SCRN [54]	.888	.885	.863	.837	.926	.927	.863	.869	.949	.916
EGNet [24]	.891	.887	.868	.841	.927	.925	.854	.852	.949	.918
DFI [55]	.892	.887	.865	.840	.924	.927	.861	.865	.951	.920
U2-Net [56]	.886	.874	.871	.847	.924	.928	.849	.844	.948	.916
GCPANet [57]	.890	.891	.860	.839	.920	.927	.853	.864	.949	.920
F3Net [58]	.902	.888	.870	.838	.927	.924	.865	.861	.953	.917
GateNet [59]	.889	.885	.862	.838	.924	.920	.858	.858	.949	.915
ITSD [60]	.895	.885	.863	.840	.927	.925	.856	.859	.952	.917
MINet [23]	.898	.884	.865	.833	.927	.925	.857	.856	.953	.919
LDF [61]	.910	.892	.874	.839	.925	.924	.872	.863	.954	.919
PSGL-Net [95]	.908	.884	.871	.833	.928	.925	.863	.060	.955	.917
Auto-MSFNet [96]	.912	.877	.869	.832	.927	.914	.866	.852	.954	.908
VST [97]	.892	.896	.861	.850	.918	.932	.844	.872	.953	.928
PFSNet [98]	.902	.892	.875	.842	.928	.930	.862	.860	.956	.924
ICON [68]	.902	.889	.870	.844	.929	.929	.861	.861	.952	.920
OLER [69]	.910	.890	.882	.845	.925	.927	.859	.857	.955	.920
Lightweight method (10M>= #Param>500K)										
HVPNet [32]	.850	.849	.839	.831	.910	.903	.830	.830	.933	.899
SAMNet [31]	.849	.849	.840	.830	.911	.907	.830	.826	.934	.898
iNAS-SOD [29]	.892	.882	.864	.839	**.927**	.923	**.863**	**.858**	.951	.917
DNTDF [30]	**.900**	**.890**	**.869**	**.841**	**.927**	**.924**	.861	**.858**	**.952**	**.920**
EDN-Lite [28]	.878	.847	.863	.823	.914	.899	.843	.820	.939	.894
Ours	.872	.852	.849	.822	.917	.911	.835	.828	.941	.902
Extremely Lightweight method (#Param)<=500K)										
CSNet [33]	.822	**.822**	.816	**.805**	**.898**	**.893**	.812	**.814**	.919	.882
CSNet* [33]	.807	.808	.802	.795	.888	.877	.811	.803	.910	.870
Ours*	**.836**	.820	**.820**	.797	.894	.888	**.820**	.810	**.926**	**.883**

Detection performance criteria. As shown in Table III and Table IV, we comprehensively evaluate all methods using four well-known evaluation metrics (mF, MAE, E_ξ, S_α). Among the lightweight methods, compared with EDN-Lite (HVPNet, SAMNet), the proposed LARNet has an average performance increase of 1% (3%, 4%), 2% (10%, 11%), similar (1%, 1%) and 1% (similar, similar) for the mF, MAE, E_ξ and S_α metrics, respectively. The metrics clearly show that compared with the other two lightweight methods, our method has greatly improved detection performance and has surpassed

Authorized licensed use limited to: UNIVERSITY OF ULSTER. Downloaded on January 19,2024 at 09:14:28 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and
content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3330082



(a) DUTS-TE (b) DUT-OMRON (c) ECSSD (d) PASCAL-S (e) HKU-IS

Fig. 4. The IOU metrics for the proposed method are compared with some state-of-the-art methods using five datasets. It is not difficult to see that our models exhibits competitive performance in (extremely) lightweight models, and even exceeds some heavyweight models in individual datasets.

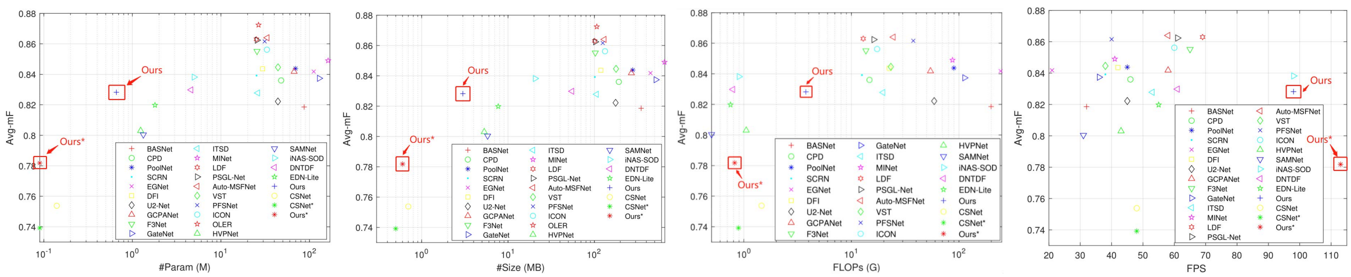


Fig. 5. The lightweight performance of the proposed method is compared with 26 state-of-the-art methods. The advantages of our proposed method are proven using four metrics: #Param, #Size, FLOPs and FPS.

some heavyweight methods in some datasets. However, compared with iNAS-SOD and DNTDF, our detection performance still has space for improvement. Considering the extremely lightweight methods, although both CSNet and CSNet* have achieved the task of SOD, their detection performance is not ideal. However, our method also greatly improves the detection performance in the extremely lightweight field and achieves optimal performance. Specifically, compared with CSNet (CSNet*), our method LARNet* has an average performance increase of 4% (6%), 9% (17%), 1% (2%) and similar (1%) in the mF, MAE, E_ξ and S_α metrics, respectively.

To evaluate the detection performance of the methods more comprehensively, as shown in Figure 4, we compare the performance of the IOU metrics for our method with other representative methods. It can be clearly seen that our (extremely) lightweight methods have competitive performance, regardless of the mean-IOU or Max-IOU metrics, but slightly inferior to iNAS-SOD and DNTDF. Similar to the detection performance metrics of the previous test, the IOU metrics for some datasets have exceeded some heavyweight methods.

Efficiency Performance Criteria. As shown in Table III, we comprehensively evaluate all methods through four commonly used evaluation metrics (#Param, #Size, FLOPs, FPS). OLER does not have complete open source code, and we cannot test it locally. The model parameters are extracted from the source paper [69]. We can clearly see that the heavyweight methods have a significant number of parameters, take up a large amount of storage, are computationally expensive and have low FPS, which poses difficultly for practical applications. Therefore, a lightweight method is needed to solve these problems. However, existing lightweight methods often have a lower FPS than heavyweight models. This may be because their unique architecture has not been optimized, and it is difficult to compete with conventional convolution with a high degree of optimization. It is worth noting that in the papers relating to HVPNet and SAMNet, their FPS reached several hundred, this is because the input batchsize is 30 as the author wants to make full use of the efficiency of GPU. However, our input batchsize is 1, which is more in line with practical applications with requirements for serial data processing. Compared with iNAS-SOD (DNTDF, EDN-Lite, HVPNet, SAMNet), the proposed LARNet reduces the model parameters and size metrics by 87% (86%, 63%, 47%, 50%) and 85% (95%, 61%, 43%, 48%), respectively. The FPS is increased by 0% (61%, 78%, 128%, 216%), and the FPS reaches approximately 98.

Among the extremely lightweight methods, CSNet and CSNet* are both powerful, but their FPS is not high (the reason is the same as HVPNet and SAMNet). Our method effectively overcomes this problem. Compared with CSNet (CSNet*), our method LARNet* reduces the model parameters, size and FLOPs metrics by 36% (the two are similar), 14% (the two are similar) and 44% (8%), respectively. The FPS is increased by 135% (135%), reaching approximately 113 FPS. Meanwhile, compared with the heavyweight method, MINet, our method LARNet (LARNet*) reduces the model parameters, size and FLOPs metrics by 99.6% (99.9%), 99.5% (99.9%) and 96% (99%), respectively. The FPS is increased by 139% (176%).

Comprehensive criteria. The above two aspects of detection performance and efficiency performance were respectively evaluated for the models. Then, we combined them to conduct a comprehensive evaluation of the methods, as shown in Figure 5. Here, Avg-mF is the average of all mF metrics across the five datasets. In the sub-figures of avg-mF vs. #Param, avg-mF vs. #Size, avg-mF vs. FLOPs and avg-mF vs. FPS, our methods show competitive performance. Although the FLOPs of LARNet is not optimal, the FPS has reached a high level. The possible reason is that different from other lightweight models, our model is built entirely on the convolutional framework optimized by PyTorch, which is more conducive

Authorized licensed use limited to: UNIVERSITY OF ULSTER. Downloaded on January 19,2024 at 09:14:28 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3330082

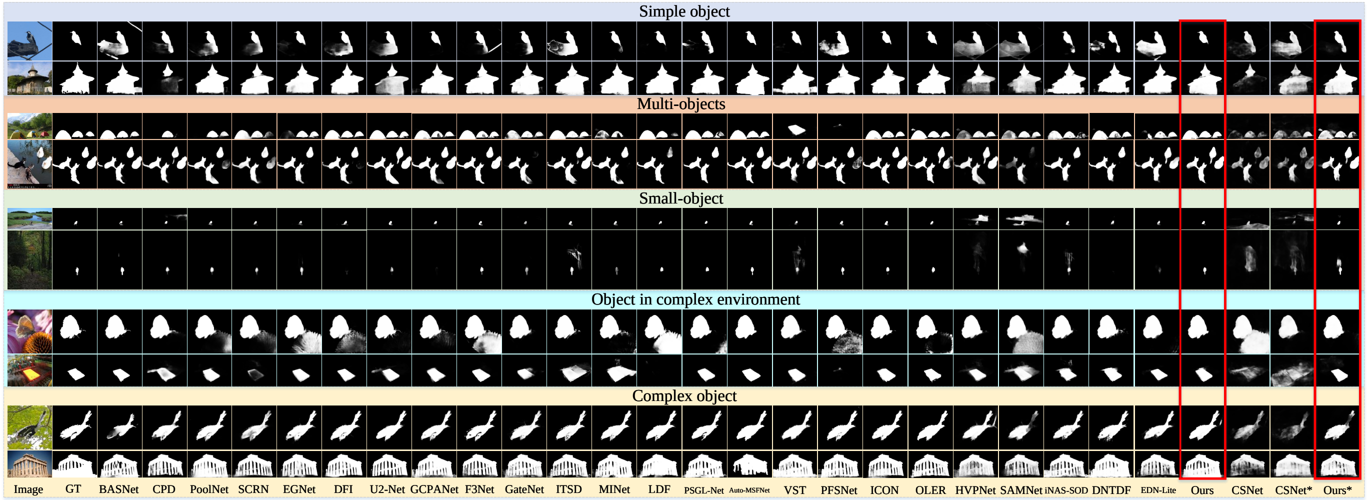


Fig. 6. Visual comparison of the proposed model with 26 state-of-the-art methods. Our methods show good performance in different scenarios, whilst meeting the needs of various computer vision tasks.

to actual deployment and application. In general, our models have the advantage in both lightweight and extremely lightweight methods, achieving significant improvement in detection performance and FPS while significantly reducing the model parameters and size. Specifically, among the lightweight methods, LARNet achieved performance improvements of 1%, 3% and 4% on the avg-mF metric, using only 63%, 47% and 50% of the parameters of EDN-Lite, HVPNet, and SAMNet, and increased FPS by 78%, 128% and 216%, respectively. Compared to iNAS-SOD (DNTDF), our model reduces the number of parameters by 87% (86%), but the avg-mF metric only decreases by 1% (the two are similar), reflecting the superiority of our model.

Among the extremely lightweight methods, although LARNet* is similar to CSNet* in terms of model parameters and size, the performance of the avg-mF metric is improved by 6% while the FPS is increased by 135%. Compared with the heavyweight model BASNet, the avg-mF metric of LARNet is increased by 1%, while the network parameters and size are reduced by 99%, FLOPs is reduced by 98%, and FPS is increased by 206%. This indicates that the detection performance of the lightweight method is close to or may even surpass that of the heavyweight method.

2) Visual Comparison: We have already demonstrated the superiority of our models at the metric level, but for visual tasks, the quality of visual saliency map generation is important. Especially when salient object detection (SOD) is one of the links of other visual tasks, whether it can show a good visual effect on the input image is particularly important. Therefore, we visually compare the saliency maps generated by the methods, as shown in Figure 6. We compare the saliency maps generated by the methods for five scenes, including simple objects (SO), multi-objects (MO), small objects (SMO), objects in a complex environment (OCE) and complex objects (CO). For the case of SO, our methods accurately locate the salient object, effectively suppressing the interference of the non-salient object(s), and the visual effect is better than many heavy-weight methods. For the case of MO, our methods are more sensitive to multi-object detection, achieve precise positioning and segmentation of salient objects, and also show more competitive visual effects than heavy-weight methods. For the case of SMO, our methods can effectively deal with the detection of small objects and accurately segment them, which is very close to the ground truth. Similarly, we still have an advantage compared with the heavy-weight models. For the case of OCE, the detection of a salient object in a complex environments is challenging, and it is often difficult to accurately locate and segment the object. However, our methods have solved this problem well and reached close to ideal results, which are on par with heavy-weight methods. For the case of CO, complex objects often have complex detailed information (such as edges), so it is particularly important to accurately recover detailed information. Our models have recovered more detailed information, and their visual effects are close to the heavy-weight methods or even surpassed.

Through the analysis of these five scenarios, it can be seen that although our models are (extremely) lightweight model, its visual effect is excellent and even comparable to some heavyweight methods, which makes it possible to embed the SOD method on a device with limited computing power, and demonstrates great potential. On the contrary, the processing effect of light-weight model on details (such as edges) still lags behind that of heavyweight model. This also shows that the lightweight model has a large space for improvement, which is a hot topic worth further study.

D. Ablation Studies

To prove the effectiveness of our proposed methods, ablation experiments are essential. Due to the similar structure of LARNet and LARNet*, we only conduct ablation experiments on LARNet. Our ablation experiments include: 1) different combinations of lightweight backbone networks; 2) different combinations of the proposed saliency backbone network LSBNet; 3) different combinations of components; 4) different

Authorized licensed use limited to: UNIVERSITY OF ULSTER. Downloaded on January 19,2024 at 09:14:28 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3330082

TABLE V

RESULTS OF THE ABLATION EXPERIMENTS. THE HIGHEST EVALUATION METRIC IS MARKED IN BOLD **RED**. W/O.: WITHOUT. OC*={16,16,24,24,32,32,64,64,64,64}. FF:FEATURE FUSION. CONV: CONVENTIONAL CONVOLUTION. ORIGINAL: MULTI-LEVEL FEATURES DIRECTLY OUTPUT BY THE BACKBONE NETWORK. S.O.:SIDE OUTPUT.

NO.	Setting	#Param (M)	#Size (MB)	FLOPs (G)	FPS	DUTS-TE		DUT-OMRON		ECSSD		PASCAL-S		HKU-IS	
						mF↑	MAE↓	mF↑	MAE↓	mF↑	MAE↓	mF↑	MAE↓	mF↑	MAE↓
	Different combinations of lightweight backbone networks														
1	MobileNet-V3	2.98	12.3	2.46	63	.788	.055	.750	.067	.908	.043	**.805**	**.080**	**.897**	**.035**
2	MobileNet-V3 w/o. pre	2.98	12.3	2.46	63	.785	.057	.746	.066	.894	.051	.793	.091	.886	.040
3	ShuffleNet-V2	0.95	4.2	**2.30**	70	.785	.055	.737	.066	.896	.047	.784	.091	.888	.040
4	ShuffleNet-V2 w/o. pre	0.95	4.2	**2.30**	70	.756	.064	.720	.073	.888	.052	.774	.095	.875	.043
5	GhostNet	2.67	11.1	**2.30**	54	.788	.056	**.751**	.066	**.909**	.042	.800	.083	.896	**.035**
6	GhostNet w/o. pre	2.67	11.1	**2.30**	54	.774	.060	.738	.071	.896	.048	.781	.092	.885	.040
7	Proposed method	**0.66**	**3.0**	3.77	**98**	**.793**	**.052**	.745	**.065**	.907	**.041**	.801	.082	.895	.036
	Different combinations of the proposed saliency backbone network LSBNet														
1	N={1,1,1,0,1,0,1,0,1,0}	0.35	1.6	3.27	**121**	.747	.063	.703	.076	.875	.054	.772	.093	.873	.043
2	N={1,1,1,1,1,1,1,1,1,1}	0.56	2.5	3.75	105	.780	.056	.738	.068	.901	.047	.794	.084	.889	.037
3	N={1,1,1,1,1,1,1,3,1,3}	0.77	3.5	3.80	91	.796	.052	**.756**	**.062**	.900	.043	**.801**	.083	.894	.037
4	N={1,2,1,2,1,2,1,3,1,3}	0.89	4.0	4.49	82	.797	.052	.753	.063	.906	.044	.799	.083	.896	.036
5	N={1,3,1,3,1,3,1,4,1,4}	1.13	4.9	5.21	74	.792	.053	.749	.066	.905	.042	.799	.083	.895	**.035**
6	OC=32	**0.29**	**1.5**	2.58	91	.777	.057	.738	.067	.891	.049	.793	.086	.884	.040
7	OC=48	0.46	2.2	3.19	90	.781	.056	.735	.068	.895	.048	.790	.088	.889	.038
8	OC=128	2.26	9.4	8.86	86	**.802**	**.050**	.753	.064	**.909**	.042	.800	**.082**	**.900**	**.035**
9	OC*	0.49	2.3	**1.70**	94	.770	.059	.733	.069	.889	.048	.782	.088	.880	.041
10	I=4	0.49	2.3	3.23	98	.786	.054	.745	.065	.899	.045	.797	**.082**	.889	.037
11	I=5	0.58	2.6	3.50	98	.789	.053	.742	.066	.903	.044	.800	.083	.891	.037
12	I=7	0.75	3.4	4.05	98	.793	.052	.750	**.062**	.902	.044	.800	**.082**	.894	.036
13	I=8	0.84	3.7	4.32	97	.791	.054	.741	.068	.906	.042	.798	.083	.894	.036
14	Proposed method	0.66	3.0	3.77	98	.793	.052	.745	.065	.907	**.041**	**.801**	**.082**	.895	.036
	Different combinations of components														
1	CGM w/o. Stage 2	0.66	3.0	3.77	115	.788	.054	.749	.064	.904	.042	.800	.083	.893	.036
2	CGM w/o. Gate 2	0.66	3.0	3.77	105	.789	.054	.740	.065	.899	.045	.796	.083	.892	.037
3	CGM w/o. Gate 1 & 2	0.64	2.9	3.60	123	.789	.053	.743	.065	.901	.046	.799	.083	.891	.037
	LSBNet CGM FF														
4	✓	**0.54**	**2.4**	**1.84**	**216**	.739	.061	.677	.079	.866	.061	.779	.090	.859	.048
5	✓ ✓	0.59	2.6	2.19	131	.757	.058	.706	.074	.882	.052	.789	.085	.876	.041
6	✓ ✓	0.62	2.7	3.42	136	.781	.055	.733	.070	.894	.047	.796	.083	.890	.037
7	CGM and FF with Conv	1.45	6.0	9.22	116	**.800**	**.050**	**.753**	**.063**	**.909**	**.040**	**.807**	**.080**	**.899**	**.034**
8	CGM with original	0.66	3.0	3.77	108	.788	.055	.743	.066	.901	.045	.805	.081	.891	.037
9	Proposed method	0.66	3.0	3.77	98	.793	.052	.745	.065	.907	.041	.801	.082	.895	.036
	Different combinations of supervision														
	BCE IOU S.O.														
1	✓	**0.66**	**3.0**	**3.77**	**98**	.741	.059	.709	.068	.887	.050	.778	.089	.868	.044
2	✓ ✓	**0.66**	**3.0**	**3.77**	**98**	.786	.054	.740	.066	.902	.043	.796	.085	.890	.038
3	✓ ✓	**0.66**	**3.0**	**3.77**	**98**	.745	.058	.710	.071	.884	.050	.779	.088	.871	.043
4	Proposed method	**0.66**	**3.0**	**3.77**	**98**	**.793**	**.052**	**.745**	**.065**	**.907**	**.041**	**.801**	**.082**	**.895**	**.036**

combinations of supervision. All the ablation experiments follow the same implementation setup to ensure fairness.

1) Ablation on existing lightweight backbone networks: There are many existing lightweight backbone networks, such as MobileNet-V3 [25], ShuffleNet-V2 [26], GhostNet [27], etc. They are all general lightweight backbone networks that can be directly applied for feature extraction in computer vision tasks through simple configuration. The design of our ablation experiment is as follows: we do not change the architecture of the existing lightweight backbone networks (MobileNet-V3 [25], ShuffleNet-V2 [26], GhostNet [27]), replacing LSBNet in LARNet with each of them respectively. Meanwhile, we loaded/unloaded the pre-trained models using ImageNet corresponding to the existing lightweight backbone network. After the multi-level feature output of the backbone network, the number of multi-level feature channels is unified to 64 through DSConv to match the requirements of LARNet for post-information processing.

As shown in Table V, we can clearly see that the method we proposed has excellent portability and can be transplanted to various existing lightweight backbone networks. CGM and the feature fusion module as a plug-and-play module also show powerful performance. Obviously, by comparing No.1 to the No.6, the pre-trained backbone network makes the model perform better than the non-pre-trained backbone network, which is consistent with our expectation.

Additionally, comparing No.1 to No.6 and No.11, in Table V, it is clear to see that our method demonstrates better overall performance than the other methods (whether or not the pre-trained model is loaded), which also illustrates that our proposed LSBNet may have more powerful performance after pre-training with ImageNet. Due to the limited laboratory resources and the fact that LARNet without pre-training has reached the best state compared with other lightweight methods (HVPNet and SAMNet), we did not pre-train the proposed LSBNet on ImageNet. According to the trend of No.1 to No.6, we have reason to believe that our backbone network will improve the performance of LARNet after pre-training. The good performance of our proposed LARNet is mainly due to the later feature processing stage that can better process the multi-level features generated by LSBNet. However, the backbone networks of MobileNet-V3 [25], ShuffleNest-V2 [26], and GhostNet [27] are relatively complex, and the redundant features generated can affect the later feature pro-

Authorized licensed use limited to: UNIVERSITY OF ULSTER. Downloaded on January 19,2024 at 09:14:28 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3330082

cessing, making the overall performance worse. The existing lightweight SOD models (such as HVPNet, SAMNet) are still competitive, which also proves the superiority and portability of our proposed method. The LSBNet we proposed generates multi-level features, less redundant features and importantly, more detailed low-level features. These features help to restore the detailed information of the output saliency map. With the help of powerful feature processing modules in the later stage, not only is the performance guaranteed, the network has also been made more lightweight. In summary, LSBNet is a lightweight saliency backbone network, which still achieves high performance without pre-training on ImageNet.

2) Ablation on the proposed backbone network LSBNet: In the previous section, we have verified the superiority of our proposed backbone network LSBNet, and now we conduct ablation experiments on LSBNet to prove that the current configuration parameters have reached the optimal effect. As shown in Table V, our ablation experiment mainly focuses on the three parameters of N, OC and I. For the ablation experiments of parameter OC, the number of multi-level feature channels is unified to 64 through DSConv to match the requirements of LARNet for post-information processing.

As shown in Table V, comparing No.1 to No.5 and No.14, we can see that as the number of layers N decreases, although the performance of the lightweight metrics is improved, the detection performance is greatly reduced. Similarly, with the increase of the number of layers N, the detection performance does not increases significantly, and the performance of the lightweight metrics decreases. This may be due to a smaller number of layers N that cannot fully extract the necessary feature information, and a larger number of layers N that introduces more redundant features. In summary, it can be seen that the number of layers N we selected provides the best overall performance. Comparing No.6 to No.9 and No.14 in Table V, we can see that as the number of output channels OC increases, the detection performance of the network also improves. This is because the network has learned more feature information. When OC=128, the improvement in network detection performance has reached a bottleneck, which may be due to the production of more redundant features, and the cost increases significantly in the lightweight metrics. In summary, we choose OC=64 to be an ideal parameter, which can demonstrate strong detection performance and lighten the network. Comparing No.10 to No.13 and No.14, we can see that as the parameter I increases, the network detection performance tends to increase while the lightweight performance decreases. Similarly, when I reaches 7, the improvement in network detection performance reaches a bottleneck, which is similar to the results when changing the layer number N. Overall, we can clearly see that the parameter selection of our saliency backbone network LSBNet is optimal, while taking into account both detection performance metrics and lightweight metrics.

3) Ablation on components: As shown in Figure 1, our method consists of three parts: LSBNet, CGM and feature fusion. In the last section, we have demonstrated the superiority of LSBNet. Therefore, the design of this ablation experiment is as follows: we keep the LSBNet configuration unchanged and then (1) discuss the advantages of the CGM module, (2) confirm the advantages of each module by loading/unloading CGM and feature fusion respectively. Since the removal of modules will cause the network to fail to operate, we adopt operations such as addition and DSConv to adapt the network so that it can still operate after the modules are removed.

As shown in Table V (Different combinations of components), compared with No.1 and No.9, we do not change the overall structure of LARNet, but only delete the Stage 2 of the CGM. Through the experiment, it can be seen that the CGM without Stage 2 makes the detection performance of LARNet decrease (especially using the DUTS-TE dataset). It shows that the introduction of multi-level features enhances the global perception ability of the input features of the CGM, thus enhancing the overall performance of the model. Compared with No.2 and No.9, we do not change the overall structure of LARNet, but only delete Gate 2 of CGM. It can be seen from the experimental data that the introduction of Gate 2 effectively improves the overall performance of LARNet. It shows that the interaction between "excitatory" and "inhibitory" neurons is more conducive to the model learning useful features. Compared with No.3 and No.9, we do not change the overall structure of LARNet, but only delete Gate 1 and Gate 2 of CGM. We can see that when CGM loses the mechanisms of "excitatory neurons" and "inhibitory neurons", the detection performance of the model decreases (especially on the ECSSD dataset), proving the rationality of the Gate mechanism. To sum up, comparing No.1, No.2, No.3 and No.9, it is easy to see that the simple fusion of multi-level features will lead to a sharp decline in network performance. This may be because of the large amount of redundant information in the multi-level features, which leads to the network failing to grasp the key information. This also illustrates the effectiveness of the interaction between the brain-inspired "excitatory" and "inhibitory" neurons. In conclusion, the above experiments prove the rational and superiority of our proposed CGM.

As shown in Table V, comparing with No.4 to No.6 and No.9, we can clearly see the superiority of each module, and the introduction of each module further improves the overall performance of the network. It is worth noting that, as seen in No.5, we add the output of CGM and directly output the saliency map, which leads to a decrease in performance, and also shows that CGM is dependent on the feature fusion module. When all the modules interact with each other, the network reaches its best state. In addition, we conducted experiments with CGM and FF under conventional convolution (No.7). Compared with No.9, the Avg-mF of No.7 was improved by 1%, but the model parameters and FLOPs increased by 120% and 145% respectively. We also input the original multi-level features from the backbone network output directly to the CGM, instead of using the output of the previous level CGM as the next level CGM input (No.8), and comparing with No.9, the overall detection performance of No.8 decreases. The above experiments have demonstrated the rational of each module design and the optimization of the combined method.

4) Ablation on supervision: Although we have built a novel SOD network, the key to determining whether the network

Authorized licensed use limited to: UNIVERSITY OF ULSTER. Downloaded on January 19,2024 at 09:14:28 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3330082

is effective lies in the reasonable use of the loss function. Therefore, the design of our ablation experiment is as follows. We prove the superiority of our proposed method by applying different supervision signals (BCE and IOU) to the final output prediction map (Sal1) and the side output prediction map (Sal2, Sal3, Sal4, Sal5 and Sal6). We then consider whether to add supervision signals to the side output prediction map.

As shown in Table V (Different combinations of components), as we only change the network supervision signal and the presence or absence of a side output, it has little or no effect on the lightweight metrics, which can be ignored. Therefore, we mainly focus on the level of its detection performance. Comparing No.1 (No.3) and the No.2 (No.4) in Table V, we can clearly see that the introduction of the IOU loss function is crucial to the improvement of network performance, which also shows that the network pays attention to the integrity of the output saliency map. Comparing No.1 (No.2) and No.3 (No.4), we can also clearly see that the introduction of the side output greatly improves the mF metric, illustrating that the introduction of more supervision signals makes the model training more stable. Our method introduces IOU loss function and side output on the basis of BCE loss function, and the overall performance of the network is significantly improved. Through this analysis, we have determined that our supervision and side output approach are reasonable and superior to other state-of-the-art approaches.

V. Conclusion

In view of the current difficulty in balancing between being lightweight, accurate and the requirement for real-time performance, we propose a novel lightweight SOD method LARNet and an extremely lightweight SOD method LARNet*. These models can be adapted for specific application requirements and are equipped with a novel (extremely) lightweight saliency backbone network, with the simplest network architecture to achieve the extraction of multi-level features, and high performance without pre-training on ImageNet. Additionally, with the introduction of the context gating module (CGM) and feature fusion module, inspired by the physiological mechanism of the human brain, the model improves the accuracy and real-time performance substantially compared with existing state-of-the-art approaches, and realizes a good balance between lightweight requirements, accuracy and real-time capability. Compared with other state-of-the-art methods, our method has advantages over (extremely) lightweight methods, it is easier to embed in resource-limited devices and achieves real-time performance. As a lightweight model, LARNet's detection performance is even better than some heavyweight methods. Through this paper, we provide new ideas for a lightweight SOD method, and further promote the development of lightweight models and the implementation of practical applications. We have also demonstrated that lightweight methods are approaching and almost surpassing the performance of heavyweight methods.

In future work, we will develop more advanced models and strategies to make the lightweight SOD models more competitive when compared with state-of-the-art heavyweight models. Additionally, we will investigate more advanced knowledge distillation methods for lightweight networks, and apply them to fields such as visual tracking [99], video object segmentation [100], etc. Furthermore, we will attempt to improve their overall performance by using lightweight SOD models as a plug-and-play module.

Acknowledgment

This work was supported by National Natural Science Foundation of China (No. 61973066), Major Science and Technology Projects of Liaoning Province (No. 2021JH1/10400049), Fundamental Research Funds for the Central Universities (N2004022).

References

[1] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," *Human Neurobiology*, vol. 4, no. 4, pp. 219–227, 1985.

[2] W. Wang, Q. Lai, H. Fu, J. Shen, H. Ling, and R. Yang, "Salient object detection in the deep learning era: An in-depth survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[3] Z. Wang, L. Du, P. Zhang, L. Li, F. Wang, S. Xu, and H. Su, "Visual attention-based target detection and discrimination for high-resolution sar images in complex scenes," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 56, no. 4, pp. 1855–1872, 2018.

[4] Z. Zhang, Z. Cui, C. Xu, Y. Yan, N. Sebe, and j. Yang, "Pattern-affinitive propagation across depth, surface normal and semantic segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 4101–4110.

[5] N. Huang, Y. Yang, D. Zhang, Q. Zhang, and J. Han, "Employing bilinear fusion and saliency prior information for rgb-d salient object detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 1651–1664, 2022.

[6] T. Zhou, D.-P. Fan, M.-M. Cheng, J. Shen, and L. Shao, "Rgb-d salient object detection: A survey," in *Computational Visual Media*, 2021, p. 37–69.

[7] K. Fu, D.-P. Fan, G.-P. Ji, Q. Zhao, J. Shen, and C. Zhu, "Siamese network for rgb-d salient object detection and beyond," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 9, pp. 5541–5559, 2022.

[8] X. Jin, K. Yi, and J. Xu, "Moadnet: Mobile asymmetric dual-stream networks for real-time and lightweight rgb-d salient object detection," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 32, no. 11, pp. 7632–7645, 2022.

[9] W. Zhou, Y. Zhu, J. Lei, R. Yang, and L. Yu, "Lsnet: Lightweight spatial boosting network for detecting salient objects in rgb-thermal images," *IEEE Transactions on Image Processing*, vol. 32, pp. 1329–1340, 2023.

[10] K. Wang, S. Ma, J. Chen, and J. Lu, "Salient bundle adjustment for visual slam," *arXiv:2012.11863*, 2020.

[11] Y. Kong, Y. Wang, and A. Li, "Spatiotemporal saliency representation learning for video action recognition," *IEEE Transactions on Multimedia*, vol. 24, pp. 1515–1528, 2022.

[12] W. Wang, J. Shen, X. Lu, S. C. H. Hoi, and H. Ling, "Paying attention to video object pattern understanding," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 7, pp. 2413–2428, 2021.

[13] W. Wang, J. Shen, J. Xie, M.-M. Cheng, H. Ling, and A. Borji, "Revisiting video saliency prediction in the deep learning era," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 43, no. 1, pp. 220–237, 2021.

[14] C. Craye, D. Filliat, and J.-F. Goudou, "Environment exploration for object-based visual saliency learning," in *IEEE International Conference on Robotics and Automation (ICRA)*, 2016, pp. 2303–2309.

[15] Z. Zhong, L. Zheng, Z. Zheng, S. Li, and Y. Yang, "Camera style adaptation for person re-identification," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018, pp. 5157–5166.

[16] T. Li, K. Zhang, S. Shen, B. Liu, Q. Liu, and Z. Li, "Image co-saliency detection and instance co-segmentation using attention graph clustering based graph convolutional network," *IEEE Transactions on Multimedia*, vol. 24, pp. 492–505, 2022.

Authorized licensed use limited to: UNIVERSITY OF ULSTER. Downloaded on January 19,2024 at 09:14:28 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3330082

[17] W. Zhou, J. Wu, J. Lei, J.-N. Hwang, and L. Yu, "Salient object detection in stereoscopic 3d images using a deep convolutional residual autoencoder," *IEEE Transactions on Multimedia*, vol. 23, pp. 3388–3399, 2021.

[18] D.-P. Fan, T. Li, Z. Lin, G.-P. Ji, D. Zhang, M.-M. Cheng, H. Fu, and J. Shen, "Re-thinking co-salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 8, pp. 4339–4354, 2022.

[19] G. Li, Z. Liu, Z. Bai, W. Lin, and H. Ling, "Lightweight salient object detection in optical remote sensing images via feature correlation," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 60, pp. 1–12, 2022.

[20] Y. Lecun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[21] E. Shelhamer, J. Long, and T. Darrell, "Fully convolutional networks for semantic segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 4, pp. 640–651, 2017.

[22] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, "An iterative and cooperative top-down and bottom-up inference network for salient object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5961–5970.

[23] Y. Pang, X. Zhao, L. Zhang, and H. Lu, "Multi-scale interactive network for salient object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 9410–9419.

[24] J.-X. Zhao, J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "Egnet: Edge guidance network for salient object detection," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 8778–8787.

[25] A. Howard, M. Sandler, B. Chen, W. Wang, L.-C. Chen, M. Tan, G. Chu, V. Vasudevan, Y. Zhu, R. Pang, H. Adam, and Q. Le, "Searching for mobilenetv3," in *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 1314–1324.

[26] N. Ma, X. Zhang, H.-T. Zheng, and J. Sun, "Shufflenet v2: Practical guidelines for efficient cnn architecture design," in *European Conference on Computer Vision (ECCV)*, 2018, pp. 122–138.

[27] K. Han, Y. Wang, Q. Tian, J. Guo, C. Xu, and C. Xu, "Ghostnet: More features from cheap operations," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 1577–1586.

[28] Y.-H. Wu, Y. Liu, L. Zhang, M.-M. Cheng, and B. Ren, "Edn: Salient object detection via extremely-downsampled network," *IEEE Transactions on Image Processing*, vol. 31, pp. 3125–3136, 2022.

[29] Y.-C. Gu, S.-H. Gao, X.-S. Cao, P. Du, S.-P. Lu, and M.-M. Cheng, "inas: Integral nas for device-aware salient object detection," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2021, pp. 4914–4924.

[30] C. Fang, H. Tian, D. Zhang, Q. Zhang, J. Han, and J. Han, "Densely nested top-down flows for salient object detection," *Science China Information Sciences*, vol. 65, p. 182103, 2022.

[31] Y. Liu, X.-Y. Zhang, J.-W. Bian, L. Zhang, and M.-M. Cheng, "Samnet: Stereoscopically attentive multi-scale network for lightweight salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 3804–3814, 2021.

[32] Y. Liu, Y.-C. Gu, X.-Y. Zhang, W. Wang, and M.-M. Cheng, "Lightweight salient object detection via hierarchical visual perception learning," *IEEE Transactions on Cybernetics*, pp. 1–11, 2020.

[33] M.-M. Cheng, S. Gao, A. Borji, Y.-Q. Tan, Z. Lin, and M. Wang, "A highly efficient model to study the semantics of salient object detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2021.

[34] M. Huang, G. Li, Z. Liu, and L. Zhu, "Lightweight distortion-aware network for salient object detection in omnidirectional images," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2023.

[35] G. R. Yang, J. D. Murray, and X.-J. Wang, "A dendritic disinhibitory circuit mechanism for pathway-specific gating," *Nature communications*, vol. 7, no. 1, pp. 1–14, 2016.

[36] W. Lee and H. Galiana, "An internally switched model of ocular tracking with prediction," *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, vol. 13, no. 2, pp. 186–193, 2005.

[37] L. Itti and C. Koch, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 11, pp. 1254–1259, 1998.

[38] Q. Lai, S. Khan, Y. Nie, H. Sun, J. Shen, and L. Shao, "Understanding more about human and machine attention in deep neural networks," *IEEE Transactions on Multimedia*, vol. 23, pp. 2086–2099, 2021.

[39] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, no. 3, pp. 569–582, 2015.

[40] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 1155–1162.

[41] Z. Jiang and L. S. Davis, "Submodular salient region detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2043–2050.

[42] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 2083–2090.

[43] L. Huo, L. Jiao, S. Wang, and S. Yang, "Object-level saliency detection with color attributes," in *Pattern Recognition*, vol. 49, 2016, pp. 162–173.

[44] A. Aksac, T. Ozyer, and R. Alhajj, "Complex networks driven salient region detection based on superpixel segmentation," in *Pattern Recognition*, vol. 66, 2017, pp. 268–279.

[45] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3166–3173.

[46] Y. Zhou, A. Mao, S. Huo, J. Lei, and S.-Y. Kung, "Salient object detection via fuzzy theory and object-level enhancement," *IEEE Transactions on Multimedia*, vol. 21, no. 1, pp. 74–85, 2019.

[47] X. Huang, Y. Zheng, J. Huang, and Y.-J. Zhang, "50 fps object-level saliency detection via maximally stable region," *IEEE Transactions on Image Processing*, vol. 29, pp. 1384–1396, 2020.

[48] Y.-Y. Zhang, S. Zhang, P. Zhang, H.-Z. Song, and X.-G. Zhang, "Local regression ranking for saliency detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 1536–1547, 2020.

[49] P. Jiang, Z. Pan, C. Tu, N. Vasconcelos, B. Chen, and J. Peng, "Super diffusion for salient object detection," *IEEE Transactions on Image Processing*, vol. 29, pp. 2903–2917, 2020.

[50] F. Qiu, S. Zhao, Y. Zhang, R. Ma, Y. Liu, Z. Wang, and S. Coleman, "Salient object detection via bilateral feature fusion and score sorting attention mechanism," in *IEEE International Conference on Multimedia and Expo (ICME)*, 2022, pp. 1–6.

[51] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "Basnet: Boundary-aware salient object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 7471–7481.

[52] Z. Wu, L. Su, and Q. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 3902–3911.

[53] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR*, 2019, pp. 3912–3921.

[54] Z. Wu, L. Su, and Q. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2019, pp. 7263–7272.

[55] J.-J. Liu, Q. Hou, and M.-M. Cheng, "Dynamic feature integration for simultaneous detection of salient object, edge and skeleton," *IEEE Transactions on Image Processing*, vol. 29, pp. 8652–8667, 2020.

[56] X. Qin, Z. Zhang, C. Huang, M. Dehghan, and M. Jagersand, "U2-net: Going deeper with nested u-structure for salient object detection," *Pattern Recognition*, vol. 106, p. 107404, 2020.

[57] Z. Chen, Q. Xu, R. Cong, and Q. Huang, "Global context-aware progressive aggregation network for salient object detection," *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 34, no. 7, pp. 10 599–10 606, 2020.

[58] J. Wei, S. Wang, and Q. Huang, "F3net: Fusion, feedback and focus for salient object detection," in *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2020, pp. 12 321–12 328.

[59] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, "Suppress and balance: A simple gated network for salient object detection," in *European Conference on Computer Vision (ECCV)*, pp. 35–51.

[60] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *IEEE/CVF*

Authorized licensed use limited to: UNIVERSITY OF ULSTER. Downloaded on January 19,2024 at 09:14:28 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3330082

Conference on Computer Vision and Pattern Recognition (CVPR), 2020, pp. 9138–9147.

[61] J. Wei, S. Wang, Z. Wu, C. Su, Q. Huang, and Q. Tian, "Label decoupling framework for salient object detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 13 022–13 031.

[62] G. Ma, C. Chen, S. Li, C. Peng, A. Hao, and H. Qin, "Salient object detection via multiple instance joint re-learning," *IEEE Transactions on Multimedia*, vol. 22, no. 2, pp. 324–336, 2020.

[63] Z. Wang, Y. Zhang, Y. Liu, S. Liu, S. Coleman, and D. Kerr, "Mfc-net : Multi-feature fusion cross neural network for salient object detection," *Image and Vision Computing*, vol. 113, p. 104243, 2021.

[64] Z. Wang, Y. Zhang, Y. Liu, Z. Wang, S. Coleman, and D. Kerr, "Tf-sod: a novel transformer framework for salient object detection," *Neural Computing and Applications*, vol. 34, p. 11789–11806, 2022.

[65] Y. Liu, Y. Zhang, Z. Wang, F. Yang, C. Qin, F. Qiu, S. Coleman, and D. Kerr, "Complementary characteristics fusion network for weakly supervised salient object detection," *Image and Vision Computing*, vol. 126, p. 104536, 2022.

[66] J. Li, Z. Pan, Q. Liu, and Z. Wang, "Stacked u-shape network with channel-wise attention for salient object detection," *IEEE Transactions on Multimedia*, vol. 23, pp. 1397–1409, 2021.

[67] B. Xu, H. Liang, R. Liang, and P. Chen, "Locate globally, segment locally: A progressive architecture with knowledge review network for salient object detection," *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, 2021.

[68] M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao, "Salient object detection via integrity learning," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2022.

[69] Z. Yao and L. Wang, "Object localization and edge refinement network for salient object detection," *Expert Systems with Applications*, vol. 213, p. 118973, 2023.

[70] Y.-f. Zhang, J. Zheng, W. Jia, W. Huang, L. Li, N. Liu, F. Li, and X. He, "Deep rgb-d saliency detection without depth," *IEEE Transactions on Multimedia*, vol. 24, pp. 755–767, 2022.

[71] Z. Yao and L. Wang, "Boundary information progressive guidance network for salient object detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 4236–4249, 2022.

[72] S. Song, Z. Miao, H. Yu, J. Fang, K. Zheng, C. Ma, and S. Wang, "Deep domain adaptation based multi-spectral salient object detection," *IEEE Transactions on Multimedia*, vol. 24, pp. 128–140, 2022.

[73] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, and E. Ding, "A mutual learning method for salient object detection with intertwined multi-supervision," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 8142–8151.

[74] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, "Inferring salient objects from human fixations," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 42, no. 8, pp. 1913–1927, 2020.

[75] Q. Lai, T. Zhou, S. Khan, H. Sun, J. Shen, and L. Shao, "Weakly supervised visual saliency prediction," *IEEE Transactions on Image Processing*, vol. 31, pp. 3111–3124, 2022.

[76] Y. Liu, D. Zhang, N. Liu, S. Xu, and J. Han, "Disentangled capsule routing for fast part-object relational saliency," *IEEE Transactions on Image Processing*, vol. 31, pp. 6719–6732, 2022.

[77] Y. Liu, Y. Zhang, Z. Wang, F. Yang, F. Qiu, S. Coleman, and D. Kerr, "A novel seminar learning framework for weakly supervised salient object detection," *Engineering Applications of Artificial Intelligence*, vol. 126, p. 106961, 2023.

[78] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 770–778.

[79] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *International Conference on Learning Representations (ICLR)*, May 2015.

[80] L. Wang, B. Lei, Q. Li, H. Su, J. Zhu, and Y. Zhong, "Triple-memory networks: A brain-inspired method for continual learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 33, no. 5, pp. 1925–1934, 2022.

[81] W. Li, H. Chen, J. Guo, Z. Zhang, and Y. Wang, "Brain-inspired multilayer perceptron with spiking neurons," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 773–783.

[82] Y. Chang, Y. Wang, J. Peng, Z. Dong, H. Li, and W. Li, "Mfs: A brain-inspired memory formation system for gan," *IEEE Transactions on Computer-Aided Design of Integrated Circuits and Systems*, vol. 41, no. 8, pp. 2598–2610, 2022.

[83] F. Zhao, Y. Zeng, and J. Bai, "Toward a brain-inspired developmental neural network based on dendritic spine dynamics," *Neural Computation*, vol. 34, no. 1, pp. 172–189, 2022.

[84] J. Li, H. Tang, and R. Yan, "A hybrid loop closure detection method based on brain-inspired models," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 14, no. 4, pp. 1532–1543, 2022.

[85] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "Cbam: Convolutional block attention module," in *European Conference on Computer Vision (ECCV)*, 9 2018, pp. 3–19.

[86] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2017, pp. 3796–3805.

[87] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3166–3173.

[88] J. Shi, Q. Yan, L. Xu, and J. Jia, "Hierarchical image saliency detection on extended cssd," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 4, pp. 717–729, 2016.

[89] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014, pp. 280–287.

[90] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015, pp. 5455–5463.

[91] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 1597–1604.

[92] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 733–740.

[93] D.-P. Fan, C. Gong, Y. Cao, B. Ren, M.-M. Cheng, and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," *International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 698–704, 2018.

[94] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, "Structure-measure: A new way to evaluate foreground maps," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 2017, pp. 4558–4567.

[95] S. Yang, W. Lin, G. Lin, Q. Jiang, and Z. Liu, "Progressive self-guided loss for salient object detection," *IEEE Transactions on Image Processing*, vol. 30, pp. 8426–8438, 2021.

[96] M. Zhang, T. Liu, Y. Piao, S. Yao, and H. Lu, "Auto-msfnet: Search multi-scale fusion network for salient object detection," in *ACM International Conference on Multimedia (ACM MM)*, 2021.

[97] N. Liu, N. Zhang, K. Wan, J. Han, and L. Shao, "Visual saliency transformer," in *IEEE/CVF International Conference on Computer Vision (ICCV)*, 04 2021.

[98] M. Ma, C. Xia, and J. Li, "Pyramidal feature shrinking for salient object detection," *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, vol. 35, no. 3, pp. 2311–2318, 2021.

[99] J. Shen, Y. Liu, X. Dong, X. Lu, F. S. Khan, and S. Hoi, "Distilled siamese networks for visual tracking," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 44, no. 12, pp. 8896–8909, 2022.

[100] Z. Zhao, S. Zhao, and J. Shen, "Real-time and light-weighted unsupervised video object segmentation network," *Pattern Recognition*, vol. 120, p. 108120, 2021.

Zhenyu Wang received the M.S. degree in electronics and communication Engineering from the Dalian Maritime University, Dalian, China, in 2019.

He is currently a Ph.D. student joint education by the Faculty of Robotics and Engineering of Northeastern University in Shenyang, China, and the Chair of Media Technology, Technical University of Munich, Germany. He has participated in several research projects and published several journal articles. His research interests include intelligent robots, computer vision.

Authorized licensed use limited to: UNIVERSITY OF ULSTER. Downloaded on January 19,2024 at 09:14:28 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Multimedia. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TMM.2023.3330082

Yunzhou Zhang received the B.S. and M.S. degrees in mechanical and electronic engineering from the National University of Defense Technology, Changsha, China, in 1997 and 2000, respectively, and the Ph.D. degree in pattern recognition and intelligent system from Northeastern University, Shenyang, China, in 2009.

He is currently a Professor with the College of Information Science and Engineering, Northeastern University. He leads the Cloud Robotics and Visual Perception Research Group. His research has been supported by funding from various sources. He has published many journal articles and conference papers. His research interests include intelligent robots, computer vision, wireless sensor networks.

Yan Liu received the B.S. degree in Mathematics and Applied Mathematics from Tonghua Normal University, Tonghua, China, in 2016, and the M.S. degree in System Theory from Northeastern University, Shenyang, China, in 2018.

She is currently a Ph.D. student at the Faculty of Robot Science and Engineering, Northeastern University, Shenyang, China. Her research interests are system control and intelligent robot.

Cao Qin received the B.S. and Ph.D degrees in automation from Northeastern University, Shenyang, China, in 2016 and 2022, respectively.

He has published several English research articles and conference papers. His research interests include visual simultaneous localization and mapping (SLAM), place recognition, and deep learning.

Sonya A. Coleman (Member, IEEE) received the B.Sc. degree (Hons.) in mathematics, statistics, and computing and the Ph.D. degree in mathematics from Ulster University, Londonderry, U.K., in 1999 and 2003, respectively.

She is currently a Professor with the School of Computing and Intelligent Systems, Ulster University, and a Cognitive Robotics Team Leader with the Intelligent Systems Research Centre. Her research has been supported by funding from various sources. She has authored or coauthored over 150 publications in robotics, image processing, and computational neuroscience.

Dermot Kerr received the B.Sc. degree (Hons.) in computing science and the Ph.D. degree in computing and engineering from Ulster University, Londonderry, U.K., in 2005 and 2008, respectively.

He is currently a Senior Lecturer with the School of Computing, Engineering and Intelligent Systems, Ulster University. His current research interests include computational intelligence, biologically inspired image processing, mathematical image processing, omnidirectional vision, and robotics.

Authorized licensed use limited to: UNIVERSITY OF ULSTER. Downloaded on January 19,2024 at 09:14:28 UTC from IEEE Xplore. Restrictions apply.
© 2023 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.