IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 March 2023, accepted 9 April 2023, date of publication 18 April 2023, date of current version 26 April 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3268114

RESEARCH ARTICLE

EC2Net: Efficient Attention-Based Cross-Context Network for Near Real-Time Salient Object Detection

NGO THIEN THU, (Member, IEEE), MD. DELOWAR HOSSAIN, AND EUI-NAM HUH, (Member, IEEE)
Department of Computer Science and Engineering, Kyung Hee University, Yongin 17104, South Korea

Corresponding author: Eui-Nam Huh (johnhuh@khu.ac.kr)

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Ministry of Science and ICT (MSIT) (No.RS-2022-00155911, Artificial Intelligence Convergence Innovation Human Resources Development (Kyung Hee University)) and grant funded by the MSIT (Artificial Intelligence Innovation Hub) under Grant 2021-0-02068.

ABSTRACT The development of salient object detection is crucial in ubiquitous applications. Existing state-of-the-art models tend to have complex designs and a significant number of parameters, prioritizing performance improvement over efficiency. Hence, there pose significant challenges to deploying them in edge devices. The intricacy in these models stems from the complicated encoder-decoder that aims to effectively generate and integrate coarse and semantic features. To address this problem, we introduced EC2Net, an efficient attention-based cross-context network for salient object detection. To start with, we introduce the shallow crossed-context aggregation (SCCA) mechanism to enhance and preserve object boundaries for shallow layers. We introduced a deep cross-context aggregation (DCCA) mechanism to enhance semantic features in deep layers. Subsequently, we introduced the dual cross-fusion module (DCFM) to efficiently merge shallow and deep features. The proposed modules complement each other, enabling EC2Net to accurately detect salient objects with reduced computational overhead. Through experiments on five standard datasets, the proposed method demonstrated competitive performance while utilizing fewer parameters, FLOPS, and memory storage than other resource-intensive models.

INDEX TERMS Attention mechanism, convolutional neural network, EC2Net, salient object detection.

I. INTRODUCTION

Salient object detection (SOD) has attracted a lot of research attention over the years as its primary aim is to mimic the human visual system by identifying the most visually prominent objects in an image. This field of research has found applications in various fields, such as segmentation, image tracking, image retrieval, image editing, and scene classification [1], [2], [3], [4], [5]. Conventional SOD models mainly rely on handcraft features to extract low-level information to derive saliency maps in top-down and bottom-up pathways [6], [7], [8], [9], [10]. Although these models can locate salient objects, the absence of global semantic information makes it challenging to detect salient objects accurately in complicated scenes. Recently, substantial advancements for SOD have been made in the deep-learning era with the help of large-scale datasets. Among of deep-learning techniques, convolution neural networks (CNN) provide end-to-end trainable neural networks, conquer the limitations in traditional models and substantially accelerate the accuracy of salient object detection. Especially, Fully Convolutional Neural Networks (FCNs) [11] can perform dense prediction on input images of any size has captured the attention of researchers. This architecture enhances the

The associate editor coordinating the review of this manuscript and approving it for publication was Hang Shen.

receptive field by stacking multiple convolution and pooling layers, which produces intricate features at multiple levels. Besides, the features aggregation mechanism also becomes an important part of SOD models. Since high-level features are abundant in semantic features for localizing objects, but they lack fine details, whereas low-level features are rich in preserving object boundaries but lack semantic information. Therefore, an efficient mechanism such as Feature Pyramid Networks [12] has been proposed to aggregate these two features for detecting salient objects. However, the advancements of these models [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], [35] come with a computational burden as they have large-size models, huge parameters, a high number of FLOPS and difficulties in deploying in edge devices. Particularly, existing models adopt encoder-decoder architecture with robust backbones, comprising multiple layers to extract both shallow and deep features, and utilizing a complex feature integration mechanism. However, this leads to a high number of parameters and computational complexity. For example, DGRL [23] has 161M parameters, requires 646MB to store the model in memory, and has a high computational cost (191.28G FLOPS). EGNet [30] has 108M parameters, 432MB memory disk. Therefore, the heaviness of these models makes them less suitable to deploy in edge devices. As the result, this necessitates the development of an efficient and lightweight method for SOD that balances accuracy and efficiency.

Several approaches have been proposed to develop a lightweight model for classification tasks (MobileNet [36], ShuffleNet [37]). However, merely applying these networks is not sufficient for salient object detection because they have few layers and are not optimized for SOD tasks. As evident from these networks, the lack of features in multi-level restricts their ability to accurately detect salient objects.

Based on our analyses, we have developed EC2Net, which comprises several modules for feature enhancement. We have introduced the shallow crossed-context aggregation mechanism (SCCA) to enrich features in shallow layers. We have introduced the deep crossed-context aggregation mechanism (DCCA) that enhances semantic features in deep layers. We have incorporated the parallel separable compact enhancement (PSE) module to capture multi-scale features. Finally, we have incorporated the dual cross-fusion mechanism (DCFM). Our proposed model efficiently complements feature representations in both shallow and deep layers by leveraging attention mechanisms. These modules work together to create a compact architecture that can accurately detect salient objects while reducing computational overhead and the number of parameters. The main contribution is as follows:

- We propose the shallow crossed-context aggregation mechanism that employs channel attention and residual spatial attention to recover and enhance low-level features for preserving boundary information.

- We present the deep crossed-context aggregation mechanism module that employs the lightweight self-attention mechanism to enhance semantic features and locate objects accurately.

- We present the dual cross-fusion module that employs criss-cross attention to merge features from shallow to deep layers that consider the inter-relationship between two features.

- With these proposed modules, we introduced EC2Net architecture for salient object detection and conducted several experiments on various datasets to demonstrate its efficiency compared to other methods.

The structure of this paper is summarized as: in section II we provide an overview of related works, then section III presents the proposed model and the details of each component. Section IV covers the experiment results on various datasets, while Section V examines the proposed modules through an ablation study. Finally, the conclusion can be found in Section V.

II. RELATED WORKS

A. SALIENT OBJECT DETECTION METHODS

In the last decades, conventional approaches have derived visual information from handcrafted features like local and global contrast [7], [8], prior cues [9], low-rank matrix recovery [10], etc. While these techniques demonstrate good performance, their lack of robust semantic information leads to lower accuracy in identifying prominent objects in complex scenes. Recently, CNN-based models have surpassed handcrafted models due to their powerful capability in extracting robust semantic features. With continued advancements, numerous models have been proposed and have demonstrated outstanding performance in salient object detection. FCN-based models are favored over other models due to their ability to generate multi-level features, which is the pioneer that has garnered significant attention in the research community [11], [18], [19], [20], [23], [24], [27]. Luo et al. [18] employed a multi-level grid to integrate both local and global information. Amulet [19] incorporated features in a multi-level manner at multiple resolutions to enhance the feature aggregation mechanism. UCF [20] addressed model uncertainty by incorporating a re-formulated drop-out mechanism after designated convolutional layers. Wang et al. [23] have introduced a recurrent mechanism to enhance the feature maps through multiple iterations. Additionally, Zhuge et al. [35] introduced a new feature aggregation mechanism to combine features with different receptive fields and enhance feature diversity. While these methods have improved performance, they are associated with high computational complexity, slow inference speed, and large model size. These limitations make it difficult to implement these large-scale methods in edge applications. Our motivation for this work stems from the necessity to devise a solution for salient object detection that achieves a balance between accuracy and efficiency.

With the growing use of deep learning in edge AI environments, there has been a surge in the development of lightweight CNN models to enhance the efficiency of salient object detection. CPD [38] used a partial decoder

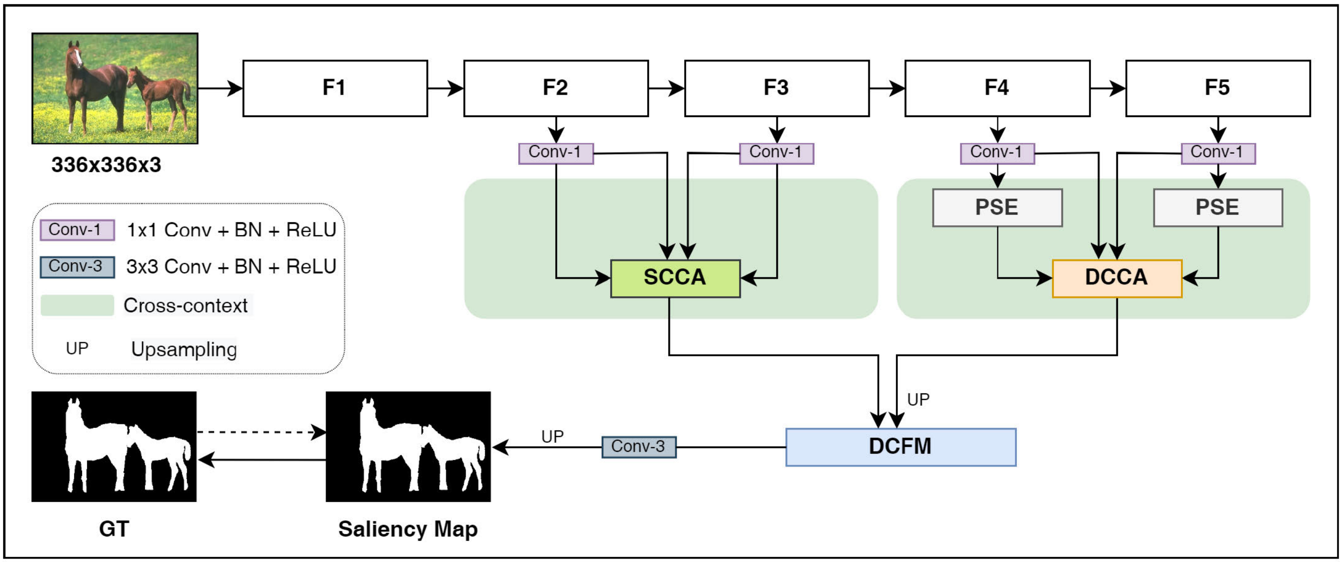

FIGURE 1. The overall architecture of EC2Net. The SCCA module efficiently enhances and aggregates features in shallow layers, PSE module captures multi-scale context while the DCCA module refines semantic context in deep layers. The dual cross fusion mechanism (DCFM) aggregates both features using an attention mechanism before generating the saliency map.

to filter out features from shallow layers and utilize the initial map to refine the features. PoolNet [33] uses pyramid pooling to merge deep semantic information with shallow features based on the Feature Pyramid Network (FPN) structure. U2Net [39] proposes a residual U-block to increase network depth and capture more contextual information without significantly increasing computational costs. HVPLNet [40] introduces a hierarchical perception module designed to mimic the structure of the primate visual cortex. SAMNet [41] developed a stereoscopic attention mechanism as a basic unit for their framework for effective multi-scale learning. They are some of the lightweight CNN models that have been developed to balance performance and efficiency in salient object detection. Although these methods have made improvements in reducing complexity, the limited architecture of the lightweight models still constrains their accuracy performance.

B. ATTENTION IN SALIENT OBJECT DETECTION

Numerous deep-learning models in computer vision have employed the attention mechanism for refining features. Multiple research studies have endeavored to enhance the performance of various baseline models by incorporating attention modules. Woo et al. combine channel and spatial attention in a sequential manner in their CBAM module [42]. Hu et al. [43] introduced the SE attention mechanism which considers channel inter-dependencies by embedding global information in each channel using its contextual information. Self-attention has been successfully recognized for its capacity to capture global information in natural language processing [44]. Huang et al. [45] devised a new attention mechanism, which acquires contextual information and reduces computational complexity for semantic segmentation. Besides, numerous attention mechanisms have been employed in salient object detection, demonstrating encouraging outcomes. SAMNet [41] utilized attention modules as fundamental building blocks for learning features at multiple scales and levels. Wang et al. [46] provide a hierarchical attention mechanism to explore multi-scale saliency. Chen et al. [26] introduce reverse attention to guide the side-out residual learning. Zhang et al. [47] introduced a bilateral attention mechanism to complement attentive features in background and foreground regions for RGB-D images. Through the use of these techniques, it has been demonstrated that the inclusion of an attention mechanism can improve a network's ability to detect significant objects. Despite these models, we employ multiple attention mechanisms to capture cross-context in both shallow and deep layers. They complement each other to enhance the features in a cross manner.

III. THE PROPOSED METHOD

A. OVERVIEW OF NETWORK ARCHITECTURE

After analyzing the limitations of current heavy models, we propose a framework that effectively leverages multiple attention mechanisms to enhance and combine features from coarse to fine, resulting in improved efficiency and reduced computational overhead. Our proposed network, called EC2Net, as illustrated in Figure 1, is built on Resnet-50 [48] as the feature extraction backbone. We have eliminated the fully connected layers and final pooling to customize the backbone network for SOD. The input image is fed into the encoder backbone for feature extraction, which produces five feature layers, denoted as F_i, where i $\in \{1,\ldots,5\}$. To reduce the number of channels, we apply 1×1 convolution on all five layers. First, we design a shallow crossed-context aggregation mechanism (SCCA) to efficiently capture and enhance cross features in shallow

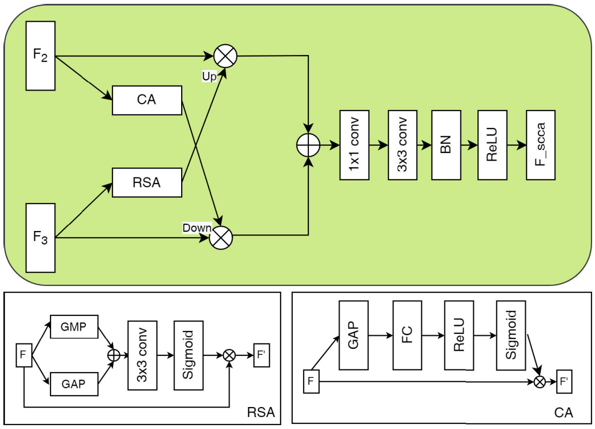

FIGURE 2. The architecture of shallow crossed-context aggregation (SCCA) mechanism.

layers. Secondly, in order to supplement multi-scale semantic information for locating salient objects accurately, we use the parallel separable compact enhancement (PSE) followed by a deep cross-context aggregation (DCCA) module to enhance high-level features. In the end, a dual cross-fusion mechanism (DCFM) is utilized to combine two branches of features for generating the final saliency map.

B. SHALLOW CROSS CONTEXT AGGREGATION MECHANISM (SCCA)

To effectively detect salient objects, it is crucial to capture detailed spatial information from shallow layers. However, downsampling operation in the encoder can significantly damage the spatial features obtained from these layers. To overcome this issue, we propose an SCCA mechanism to recover and aggregate discriminative features in shallow layers effectively. As depicted in Figure 2, SCCA comprises two attention modules followed by a cross-aggregation operation.

Due to the fact that prominent objects are typically situated in the foreground areas, indiscriminately treating all spatial pixels as equal can result in incorrect detections. Moreover, different channels provide different contextual information, so treating all channels in the same manner, can lead to inaccurate outcomes. Hence, to extract the optimal features in shallow layers, we use cross-aggregation operation on channel-wise attention (CA) and residual spatial-wise attention (RSA) respectively.

First, the CA module calculates a channel-wise attention map by considering the inter-correlation among different channels [43]. We use the second layer F$_2$ to compute channel attention features F$_{ca}$ as below formula:

$$c_a = \sigma(FC(GAP(F_2))) \tag{1}$$
$$F_{ca} = F_3 \times Sigmoid(c_a) \tag{2}$$

where GAP is the global average pooling, FC performs the dimensional reduction on the output of GAP, and σ is the ReLU activation function. Next, we use the third layer F$_3$ to calculate the spatial-wise attention features F$_{rsa}$. RSA module helps highlight important features at each spatial location. Particularly, the process of RSA is followed as:

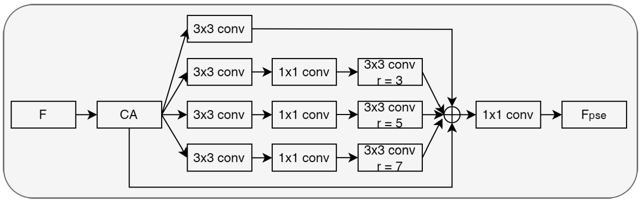

FIGURE 3. The architecture of parallel separable compact enhancement (PSE) mechanism.

$$s_a = conv3(GAP(F_3) + GMP(F_3)) \tag{3}$$
$$F_{rsa} = conv3(F \times Sigmoid(s_a)) + F_3 \tag{4}$$

where GAP and GMP denote the global average pooling and global max pooling respectively, and conv3 denotes 3 × 3 convolution, followed by Batch Normalization (BN) and ReLU activation function.

Finally, the SCCA module aggregates the above attentive features to capture cross-context in shallow layers. Since low-level features carry more abstract information and mid-level features contain less semantic but more detailed information. We combine two feature maps in a cross manner to complement feature representations in low and mid features, resulting in enhanced quality of the output features. The two outputs from two attention modules are denoted as F$_{ca}$ $\in \mathbb{R}^{H\times W\times C}$, and F$_{rsa}$ $\in \mathbb{R}^{\frac{H}{2}\times\frac{W}{2}\times 1}$. To account for the difference in spatial resolutions between F$_{ca}$ and F$_{rsa}$, we apply upsampling and downsampling operations before the crossing process. Specifically, we upsample F$_{rsa}$ to the same size as F$_2$, resulting in dimensions of $\mathbb{R}^{H\times W\times 1}$. Similarly, we downsample F$_{ca}$ to the size of F$_3$, resulting in dimensions of $R^{\frac{H}{2}\times\frac{W}{2}\times C}$. We then multiply these features in a cross manner as depicted in Fig.2. Then, we concatenate the output from two branches and 1 × 1 convolution to reduce the number of channels, before using another 3 × 3 convolution followed by BN and ReLU functions to refine the features. The crossing aggregation process is summarized below:

$$F_{sca1} = upsample(F_rsa) \times F_2 \tag{5}$$
$$F_{sca2} = downsample(F_ca) \times F_3 \tag{6}$$
$$F_{scca} = \sigma(BN(conv3(conv1(F_sca1 + upsample(F_sca2))))) \tag{7}$$

By using the SCCA module, we can obtain a richer feature representation while preserving fine-grained information in shallow layers. Moreover, the attention-based concept helps reduce noise and recover lost information more effectively. The output of this module is denoted as F$_{scca}$.

C. PARALLEL SEPARABLE COMPACT ENHANCEMENT (PSE) MECHANISM

Salient objects can manifest in various sizes and shapes in real-world scenarios. Previous methods have utilized pooling layers to obtain features in a multi-scale manner, which may not effectively handle complicated structures with accuracy.

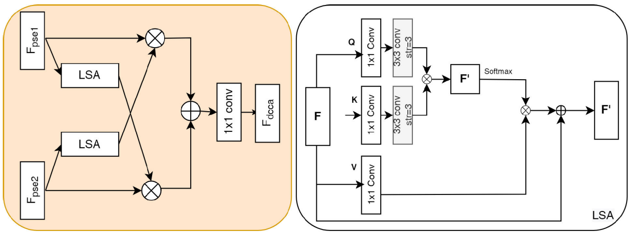

FIGURE 4. **The architecture of deep cross-context aggregation (DCCA) mechanism.**

To address this, we introduced the parallel separable compact enhancement module (PSE) that efficiently emphasizes important semantic features and captures multi-scale context. As depicted in Fig.3, the PSE module is motivated by the structure of ASPP+ [49]. Since different channels provide distinct contextual information, we employ channel attention to generate an attention map by analyzing the intercorrelation among different channels, as in equations (1) and (2). The enhanced features are then separated into four parallel branches to capture multi-scale features. We employ dilated separable convolution in this module to leverage its advantage of both dilated and separable convolutions. Dilated convolution extends the receptive field block, allowing for the capture of multi-scale semantic features, while separable convolution [50] factorizes the convolution into two smaller convolutions, resulting in a more efficient operation with fewer computations.

The first branch of PSE module contains a 3 × 3 depthwise separable convolution, the last three branches contain a 3 × 3 depthwise separable convolution, a 1 × 1 convolution, and a 3 × 3 dilated separable convolution with different rates. We insert 1 × 1 convolution in between two 3 × 3 convolutions is to reduce the number of input channels to the next 3 × 3 convolution layer. After each convolutional layer, we apply BN and ReLU functions. We use three 3 × 3 dilated separable convolutional layers with a dilation rate of 3,5,7, respectively. This helps to expand the receptive field block and capture multi-scale context without increasing the kernel size. The three outputs are concatenated with the input features before being passed through a 1 × 1 convolution layer to obtain the final features, resulting in the multi-scale features denoted as F$_{p}se$.

D. DEEP CROSS-CONTEXT AGGREGATION (DCCA) MECHANISM

After capturing multi-scale features from the PSE module, we possess valuable semantic information using DCCA module to capture the cross-context in deep layers. This help enhance the global context that plays a vital role in accurately detecting significant objects. First, we introduce a lightweight self-attention module to filter rich semantic features in two deep layers separately. Later, we generate the cross-context in high-level features by aggregating their output in cross-manner. The visualization of DCCA is depicted in Figure. 4.

In several vision-related tasks, researchers have found self-attention to be a successful approach for capturing long-range dependencies. Nevertheless, the computational complexity of self-attention can be challenging, particularly in lightweight models, as it computes attention maps for all pairwise positions. To enable efficient salient object detection, we have integrated a lightweight self-attention module that can extract global information in high-level features and overcomes the challenge. The traditional self-attention approach, described in Vaswani et al.'s work, operates on an input feature image X ∈ $\mathbb{R}^{H\times W\times C}$, where H, W, and C denote height, width, and number of channels, respectively. In this approach, 1 × 1 convolutions are employed to transform the feature map X into three sub-spaces, designated as Q, K, V ∈ $\mathbb{R}^{H\times W\times C'}$, where C' is the reduced number of channels compared to C. Subsequently, a weight matrix is obtained by calculating the correspondence between all positions in Q and K, resulting in a similarity matrix A ∈ $\mathbb{R}^{HW\times HW}$. The weight matrix A is then normalized using a soft-max function and utilized to linearly combine the values in the V sub-space, which generates the output of the self-attention layer. This mathematical process can be expressed as follows:

$$H_{sa} = Softmax(Q^T K)V \tag{8}$$

In the above equation, A = $\sigma(Q^T K)$, and A ∈ $\mathbb{R}^{HW\times HW}$. To obtain the self-attention map, the features V and the weight matrix A are multiplied and summed. The time complexity of calculating the weight matrix A using conventional self-attention is $\mathcal{O}(C'H^2W^2)$, with the dot products between every pair of input representatives being the dominant factor. To reduce the complexity of computing A, after performing 1 × 1 convolution to transform input feature to sub-spaces, we use a 3 × 3 convolution with strides 3 to perform spatial dimensional reduction on K and V vectors. The number of pixels in the weight matrix A is reduced from HxW to P, where P is smaller than HxW. Given the input from the PSE module, the mathematical equation for the lightweight version of self-attention can be presented as follows:

$$H_{lsa} = \sigma(Q^T d(K))d(V) \tag{9}$$

$$F_{lsa} = H_{lsa} + F_{pse} \tag{10}$$

where d denotes the 3 × 3 convolution with stride 3. The attentive features H_{lsa} are then connected to the input itself, followed by the loop-edge embedding concept for intra-attention [51]. In this way, the time complexity of the total process is scaled down from $\mathcal{O}(C'H^2W^2)$ to $\mathcal{O}(C'HWP)$. This process facilities the use of self-attention on resource-constrained devices.

After processing self-attention on the output of PSE module, we got two outputs F$_{lsa1}$ and F$_{ls2}$ respectively. To produce the cross high-level features denoted as F$_{dcca}$, we multiply two features F$_{lsa1}$ and F$_{lsa2}$ with the input features F$_4$ and F$_5$ in a cross manner to complement semantic information in deep layers. The concatenated features are processed through 1 × 1 convolution to generate the output F_{dcca}. The whole process is described below:

$$F_{dcca} = conv1(F_{lsa1} * F_5 + F_{la2} * F_4) \tag{11}$$

E. DUAL CROSS FUSION MODULE (DCFM)

Shallow layers offer local details that preserve object boundaries and fine-grained information, while deep layers provide more semantic features to extract salient objects. Although shallow layers provide less semantic information, they are still valuable in conjunction with the high-level information provided by deep layers. Consequently, feature fusion has become a crucial component in integrating different levels of features in SOD. A simple combination operation such as element-wise or concatenation can ignore the relationship between these two features and introduces the loss of semantic features and the degradation of object boundaries. By considering this issue, we designed a dual cross-fusion mechanism (DCFM) to employ the inter-correlation between shallow and deep features. We employ a criss-cross attention [45] mechanism to aggregate both features by reweighting the fusion process.

As shown in Figure 5, we modify the original version of criss-cross attention by using two input features. Given feature maps F$_{scca}$ $\in \mathbb{R}^{H\times W\times C}$ and F$_{dcca}$ $\in \mathbb{R}^{\frac{H}{4}\times\frac{W}{4}\times C}$, we upsample F$_{dca}$ to the same size as F$_{scca}$ before sending to DCFM module. Then we apply 1 × 1 convolutions to calculate *key* (K) and *value* (V) F$_{dcca}$ and calculate *query* (Q) from low-level features F$_{scca}$. Where query Q and Key V $\in \mathbb{R}^{H\times W\times C'}$ and C' has fewer channels. After generating Q and K, we compute the feature map Λ, where Λ measures the correlation energy between the two features in horizontal and vertical directions. First, we calculate Q$_z$ $\in \mathbb{R}^{C'}$ at each spatial position z in features Q and the set ω_z by generating features from the *key* vector K that has a similar position with position z, where $\omega_{i,z} \in \mathbb{R}^{C'}$ is the components of ω_z. The affinity operation is calculated as follows

$$\Lambda_{iz} = Q_z(\omega_i, z)^T \tag{12}$$

where $\Lambda_{i,z}$ refers the correlation energy between Q$_z$ and $\omega_{i,z}$. Then we apply the Softmax function to the energy matrix Λ over the horizontal and vertical dimensions to generate the attention map D. Similarly, we obtain vector V$_z$ and the set υ_z from features map V. The final feature attention is finally obtained as below equation:

$$F'_z = \sum_{i\in|\upsilon_z|} D_i, z\upsilon_i, z \tag{13}$$

where F'$_z$ $\in \mathbb{R}^{H\times W\times C}$ is the features vector at position z in the output feature F'. D$_{i,z}$ is the scalar value at channel i and position z in A. The attention features are added to the input features F$_{dcca}$ before sending to 3 × 3 convolution to refine the feature representation, as below equation

$$F_{dcfm} = conv3(F'_z + F_{dcca}) \tag{14}$$

F. LOSS FUNCTION

A loss function is a crucial element in SOD, as it assesses how well the algorithm performs on the experimental dataset. Binary cross-entropy (BCE) [52], also called log loss, is a frequently used loss function in SOD tasks. It measures the discrepancy between the predicted output and the true binary label (0 or 1) for each sample in a dataset. In this work, we use this binary cross-entropy formula that can be represented as:

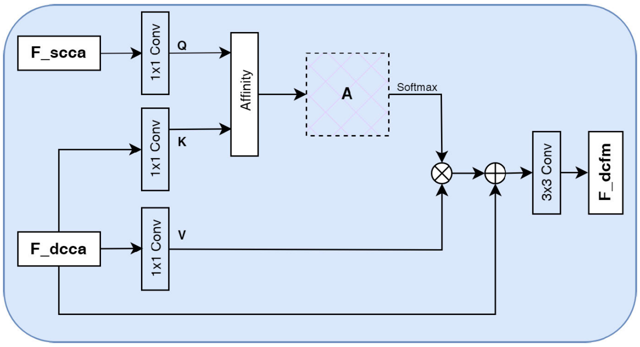

FIGURE 5. The architecture of dual cross fusion mechanism (DCFM).

$$L_{bce} = -\sum_{i=1}^{H}\sum_{j=1}^{W} G(i,j)\log(S(i,j)) + (1 - G(i,j)\log(1 - S(i,j) \tag{15}$$

IV. EXPERIMENTS AND RESULTS

A. DATASETS

We performed a comprehensive performance evaluation of our proposed approach on five prominent benchmark datasets: ECSSD [53], DUT-OMRON [54], HKU-IS [55], SOD [56], and DUTS [57].

The ECSSD dataset, an Extended Complex Scene Saliency Dataset, consists of 1000 images that showcase complex textures and structures similar to real-world scenarios. The SOD dataset comprises 300 images with their corresponding labels. The DUTS dataset has two subsets: DUTS-TR, a training set with 10,553 images, and DUTS-TE, a testing set with 5019 images, both of which exhibit complicated scenes. The DUT-OMRON dataset comprises 5019 images with varying content and complex backgrounds. Lastly, the HKU-IS dataset includes 4447 images featuring multiple salient objects under low-contrast conditions.

B. IMPLEMENTATION DETAILS

We implemented EC2Net using Pytorch 3.8 and conducted experiments on Ubuntu PC, using Geforce GTX 1080 GPU. EC2Net is trained on the DUTS-TR dataset using the Adam optimizer, with parameters set as a weight decay of 5e-3, learning rate of 5e-5. To avoid model overfitting, we employ data augmentation strategies such as random cropping, flipping, and normalization. All images are scaled to the size 336 × 336 during the training and testing process. Our network converged in 45 epochs, which took approximately 9 hours to train from scratch.

C. EVALUATION METHODOLOGY

We assess the performance of our approach from two perspectives: performance and efficiency. Given both N

TABLE 1. The performance comparison of existing methods and state-of-the-art methods in terms of F-measure, Mean Absolute Error(MAE), and S-measure. ↑ (↓) indicates higher (lower) values are better.

Methods	ECSSD			SOD			DUTS-TE			DUT-OMRON			HKU-IS		
	F↑	M↓	S↑	F↑	M↓	S↑	F↑	M↓	S↑	F↑	M↓	S↑	F↑	M↓	S↑
RFCN [15]	0.896	0.097	0.856	0.802	0.161	0.716	0.782	0.089	0.793	0.738	0.095	0.774	0.892	0.080	0.858
DHS [16]	0.905	0.062	0.880	0.822	0.128	0.746	0.815	0.065	0.820	-	-	-	0.892	0.052	0.870
DCL [17]	0.895	0.080	0.869	0.831	0.131	0.756	0.785	0.082	0.803	0.733	0.095	0.762	0.892	0.063	0.871
NLDF [18]	0.903	0.065	0.875	0.837	0.123	0.753	0.816	0.065	0.815	0.753	0.079	0.817	0.902	0.048	0.878
Amulet [19]	0.916	0.059	0.894	0.795	0.144	0.750	0.751	0.084	0.804	0.743	0.097	0.780	0.899	0.050	0.886
UCF [20]	0.903	0.069	0.883	0.805	0.148	0.759	0.773	0.112	0.782	0.730	0.120	0.759	0.888	0.061	0.874
SRM [21]	0.917	0.056	0.895	0.840	0.126	0.739	0.826	0.059	0.836	0.745	0.069	0.797	0.906	0.046	0.886
DSS [22]	0.906	0.064	0.873	0.842	0.122	0.746	0.813	0.065	0.824	0.760	0.074	0.790	0.900	0.050	0.878
DGRL [23]	0.922	0.041	0.902	0.843	0.103	0.768	0.828	0.049	0.842	0.774	0.062	0.805	0.910	0.036	0.894
RAS [26]	0.901	0.055	0.892	0.847	0.123	0.761	0.802	0.059	0.838	0.799	0.058	0.825	0.927	0.036	0.906
PiCANet [27]	0.923	0.049	0.909	0.836	0.102	0.787	0.837	0.054	0.860	0.766	0.068	0.826	0.916	0.042	0.905
BASNet [29]	0.938	0.040	0.910	0.849	0.112	0.766	0.872	0.040	0.878	0.803	0.059	0.832	0.932	0.031	0.915
EGNet [30]	0.938	0.044	0.913	0.859	0.110	0.781	0.870	0.044	0.878	0.794	0.056	0.836	0.928	0.034	0.912
CKT [31]	0.907	0.057	0.891	0.819	0.122	0.757	0.811	0.062	0.831	0.759	0.072	0.799	0.898	0.046	0.889
AFNet [32]	0.930	0.045	0.907	0.848	0.108	0.773	0.857	0.046	0.867	0.784	0.057	0.907	0.921	0.036	0.905
EC2Net	0.932	0.045	0.910	0.836	0.111	0.765	0.858	0.046	0.864	0.784	0.059	0.813	0.919	0.036	0.904

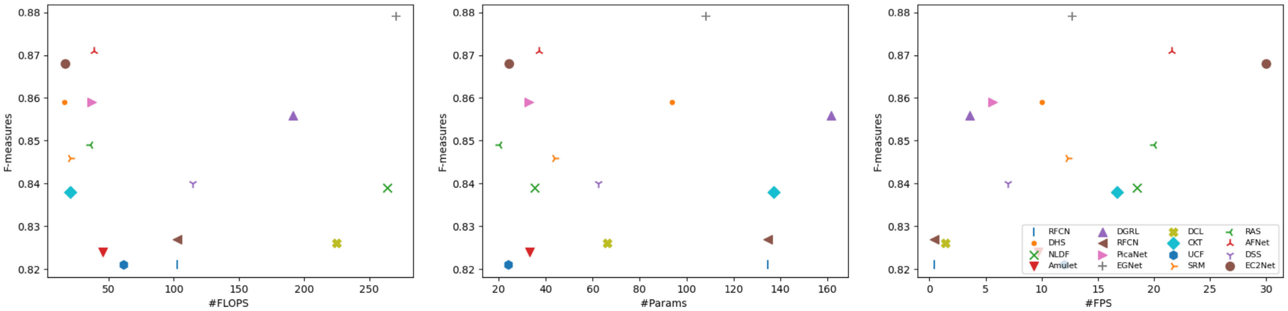

FIGURE 6. The visualization of the efficiency comparison, shows the trade-off between accuracy and efficiency. F-measure is averaged on five datasets. Our proposed method is marked as a big brown circle.

TABLE 2. The efficiency comparison of State-of-the-art methods and EC2Net on four metrics #Params, FLOPS, FPS and Size (MB).

Methods	Params (M)	FLOPS (G)	FPS	Size (MB)
RFCN [15]	134.69	102.8	0.4	-
DHS [16]	93.76	15.8	10.0	-
DCL [17]	66.24	224.9	1.4	-
NLDF [18]	35.49	263.9	18.5	428
Amulet [19]	33.15	45.3	9.7	132
UCF [20]	23.98	61.4	12	117
SRM [21]	43.74	20.3	12.3	189
DSS [22]	62.23	114.6	7.0	237
DGRL [23]	126.35	24	3.6	646
RAS [26]	20.13	35.6	20	81
PiCANet [27]	32.85	37.1	5.6	197.2
BASNet [29]	55.90	121.6	12.5	348
EGNet [30]	108.07	270.8	12.7	432
CKT [31]	137.03	20.5	16.7	-
AFNet [32]	37.11	38.4	21.6	143
EC2Net	24.38	14.19	27	89

(predicted saliency map) and G (ground truth map) are within the range of [0,1], the Mean Absolute Error (MAE) [7] value is calculated using the following equation::

$$MAE = \frac{1}{Width \times Height} \sum_{i=1}^{Width} \sum_{j=1}^{Height} |N(i,j) - G(i,j)| \quad (16)$$

A smaller MAE value indicates better performance. The F-measure [58] is a measure to combine precision (p) and recall (r) into a single score. The MaxF score is calculated from the maximum F-measure value obtained across all images in each dataset. A higher F-measure value indicates better performance. The definitions for F-measure are provided below:

$$F_\beta = \frac{(1+\beta^2)rp}{\beta^2 * p + r} \quad (17)$$

where β is set as 0.3 to maximize the precision. Finally, S-measure [59] assesses structural information that cannot be captured by pixel-based metrics such as precision and recall. S-measure is defined as:

$$S_{measure} = \lambda \times S_0 + (1-\lambda) \times S_r \quad (18)$$

where S_0 measures the objectness of the predicted saliency map and S_r measures the region similarity. We set $\lambda = 0.5$. A higher value of S-measure indicates that the model has better structural performance.

Besides, we evaluate the effectiveness of EC2Net in comparison to other models based on several metrics, including the number of parameters (#Params), FLOPS, inference speed (FPS), and size (MB). FLOPS represents a measure of the computational complexity of the model, where

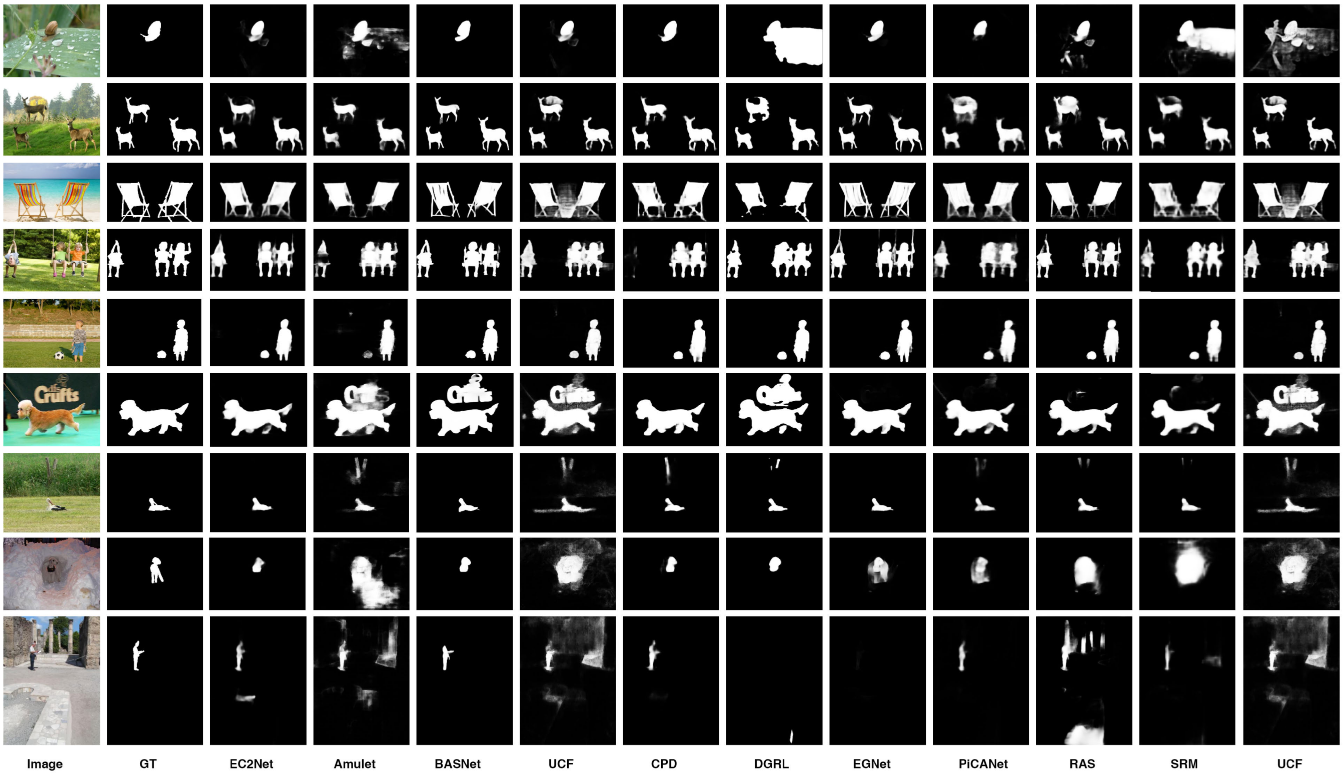

FIGURE 7. The qualitative comparison between EC2Net and state-of-the-art SOD methods. Our proposed network has good accuracy and fine boundary compared to others.

a higher FLOPS value corresponds to a more computationally intensive model. Size refers to the amount of memory required to store the network, while FPS denotes the number of frames processed by the model per second. #Params are measured in Mega (M), whereas FLOPS are expressed in Giga (G).

D. COMPARED WITH STATE-OF-THE-ART SOD MODELS

In this section, we compare the performance of EC2Net with 14 state-of-the-art methods, we select the below models for doing experiments. RFCN [15], DHS [16], DCL [17], NLDF [18], Amulet [19], UCF [20], SRM [21], DSS [22], DGRL [23], RAS [26], PiCANet [27], BASNet [29], EGNet [30], CKT [31], AFNet [32]. We collect the pre-computed saliency maps from the author's project website. As shown in Table 1, we perform the quantitative analysis on five widely used datasets based on three metrics: maximum F-measure, average MAE, and S-measure. The findings indicate that EC2Net performs at a similar level of accuracy across different popular datasets for all three metrics. Particularly, the proposed approach achieves comparable performance to other heavyweight state-of-the-art models such as EGNet, DGRL, and RFCN, but with significantly fewer parameters and FLOPS. Specifically, our method achieves the second-best F-measure score on the ECSSD dataset with a score of 0.932, compared to CKT's score of 0.907 EC2Net requires only 17.5% of CKT's parameters and has twice FPS. The proposed model also beats DGRL at the score of 0.922, while using only 14.9% of DGRL's parameters. EC2Net significantly outperforms RFCN with a score of 0.896 while only taking 17.8% number of the parameters. These results suggest that EC2Net efficiently trades off between performance and computational complexity.

In addition, we present an efficiency comparison between the proposed approach and existing models in Figure 6. Specifically, we plot the F-measure against the number of parameters, FLOPS, and FPS. For each sub-figure, we compute the average F-measure score across the five datasets. In the sub-figures depicting F-measure versus Parameters and FLOPS, our method, EC2Net, is positioned in the top-left quadrant of the chart, indicating a favorable trade-off between accuracy and efficiency.

In order to qualitatively compare EC2Net with other methods, we present a selection of representative scenes in Figure 7 to highlight the advantages of our approach over other methods. The saliency maps were chosen to represent a variety of scenarios, including images with one or multiple objects, large and small objects, and low-contrast images. EC2Net generates saliency maps that exhibit distinct and precise boundaries, closely resembling the ground truth.

E. ABLATION STUDY

To evaluate the significance of each module in EC2Net and their efficiency, we conducted ablation experiments on

TABLE 3. The effectiveness of each module in terms of F-measure, MAE, and S-measure on ECSSD, DUTS-TE, DUT-OMRON, and HKUIS datasets.

Methods	ECSSD			DUTS-TE		
	F↑	M↓	S↑	F↑	M↓	S↑
+ SCCA only	0.848	0.105	0.830	0.723	0.109	0.758
+ DCCA only	0.916	0.054	0.900	0.823	0.057	0.844
+ w/o PSE	0.922	0.052	0.903	0.837	0.054	0.852
+ w/o LSA	0.925	0.049	0.907	0.842	0.052	0.858

Methods	DUT-OMRON			HKUIS		
	F↑	M↓	S↑	F↑	M↓	S↑
+ SCCA only	0.665	0.129	0.717	0.845	0.084	0.838
+ DCCA only	0.772	0.068	0.814	0.901	0.050	0.888
+ w/o PSE	0.785	0.065	0.820	0.906	0.047	0.893
+ w/o LSA	0.781	0.069	0.817	0.911	0.044	0.898

ECSSD, DUTS-TE, DUT-OMRON, and HKUIS datasets using three metrics as presented in Table 3. We evaluated the performance of EC2Net with each module SCCA, DCCA, DCCA without LSA, and DCCA without PSE. Besides, we test inference speed with the input size 224 × 224 and achieved a real-time FPS of 35 on overall five datasets. The results show that SCCA and DCCA modules contribute to enhancing features in both shallow and deep layers, while PSE and LSA complement each other to enrich the multi-scale semantic features.

V. CONCLUSION

In this paper, we focus on balancing performance and efficiency, despite the prior research in this area. We propose three new modules to enhance and combine cross-level features. The first module, attention-based SCCA, works with shallow layers to complement cross-level features. SCCA effectively complement features low-level features and mid-level features. The second module, DCCA, captures cross-semantic features using lightweight self-attention. Hence, DCCA enhances semantic features to locate objects more accurately. Lastly, the DCFM module aggregates feature from both shallow and deep layers using a modified criss-cross attention mechanism. Our experiments on five datasets show that our approach achieves comparable performance to heavier models while using fewer parameters, FLOPS, and overall model size.

REFERENCES

[1] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 817–824.

[2] C. Ma, Z. Miao, X.-P. Zhang, and M. Li, "A saliency prior context model for real-time object tracking," *IEEE Trans. Multimedia*, vol. 19, no. 11, pp. 2415–2424, Nov. 2017.

[3] H. Wang, Z. Li, Y. Li, B. B. Gupta, and C. Choi, "Visual saliency guided complex image retrieval," *Pattern Recognit. Lett.*, vol. 130, pp. 64–72, Feb. 2020.

[4] A. Hagiwara, A. Sugimoto, and K. Kawamoto, "Saliency-based image editing for guiding visual attention," in *Proc. 1st Int. workshop Pervasive Eye Tracking Mobile Eye-Based Interact.*, Sep. 2011, pp. 43–48.

[5] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 24, no. 5, pp. 769–779, May 2014.

[6] S. Goferman, L. Zelnik-Manor, and A. Tal, "Context-aware saliency detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 10, pp. 1915–1926, Oct. 2012.

[7] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.

[8] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, Mar. 2015.

[9] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2012, pp. 29–42.

[10] H. Peng, B. Li, H. Ling, W. Hu, W. Xiong, and S. J. Maybank, "Salient object detection via structured matrix decomposition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 4, pp. 818–832, Apr. 2017.

[11] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[12] T.-Y. Lin, P. Dollar, R. Girshick, K. He, B. Hariharan, and S. Belongie, "Feature pyramid networks for object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 2117–2125.

[13] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 1265–1274.

[14] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "DeepSaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3919–3930, Aug. 2016.

[15] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2016, pp. 825–841.

[16] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 678–686.

[17] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 478–487.

[18] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 6609–6617.

[19] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 202–211.

[20] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 212–221.

[21] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4019–4028.

[22] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 3203–3212.

[23] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, "Detect globally, refine locally: A novel approach to saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3127–3135.

[24] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 714–722.

[25] J. Wei and B. Zhong, "Saliency detection using fully convolutional network," in *Proc. Chin. Autom. Congr. (CAC)*, Nov. 2018, pp. 3902–3907.

[26] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 234–250.

[27] N. Liu, J. Han, and M.-H. Yang, "PiCANet: Learning pixel-wise contextual attention for saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3089–3098.

[28] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 8150–8159.

[29] X. Qin, Z. Zhang, C. Huang, C. Gao, M. Dehghan, and M. Jagersand, "BASNet: Boundary-aware salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 7479–7489.

[30] J. Zhao, J.-J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, ''EGNet: Edge guidance network for salient object detection,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 8779–8788.

[31] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, ''Contour knowledge transfer for salient object detection,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 355–370.

[32] M. Feng, H. Lu, and E. Ding, ''Attentive feedback network for boundary-aware salient object detection,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1623–1632.

[33] 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 *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2019, pp. 3917–3926.

[34] Y. Pang, X. Zhao, L. Zhang, and H. Lu, ''Multi-scale interactive network for salient object detection,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 9413–9422.

[35] M. Zhuge, D.-P. Fan, N. Liu, D. Zhang, D. Xu, and L. Shao, ''Salient object detection via integrity learning,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 3, pp. 3738–3752, Mar. 2023.

[36] A. G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, and H. Adam, ''MobileNets: Efficient convolutional neural networks for mobile vision applications,'' 2017, *arXiv:1704.04861*.

[37] X. Zhang, X. Zhou, M. Lin, and J. Sun, ''ShuffleNet: An extremely efficient convolutional neural network for mobile devices,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 6848–6856.

[38] Z. Wu, L. Su, and Q. Huang, ''Cascaded partial decoder for fast and accurate salient object detection,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3907–3916.

[39] X. Qin, Z. Zhang, C. Huang, M. Dehghan, O. R. Zaiane, and M. Jagersand, ''U2-Net: Going deeper with nested u-structure for salient object detection,'' *Pattern Recognit.*, vol. 106, Oct. 2020, Art. no. 107404.

[40] Y. Liu, Y.-C. Gu, X.-Y. Zhang, W. Wang, and M.-M. Cheng, ''Lightweight salient object detection via hierarchical visual perception learning,'' *IEEE Trans. Cybern.*, vol. 51, no. 9, pp. 4439–4449, Sep. 2020.

[41] 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 Trans. Image Process.*, vol. 30, pp. 3804–3814, 2021.

[42] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, ''CBAM: Convolutional block attention module,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 3–19.

[43] J. Hu, L. Shen, and G. Sun, ''Squeeze-and-excitation networks,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 7132–7141.

[44] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, ''Attention is all you need,'' in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 1–15.

[45] Z. Huang, X. Wang, L. Huang, C. Huang, Y. Wei, and W. Liu, ''CCNet: Criss-cross attention for semantic segmentation,'' in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 603–612.

[46] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, ''Salient object detection with pyramid attention and salient edges,'' in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1448–1457.

[47] Z. Zhang, Z. Lin, J. Xu, W.-D. Jin, S.-P. Lu, and D.-P. Fan, ''Bilateral attention network for RGB-D salient object detection,'' *IEEE Trans. Image Process.*, vol. 30, pp. 1949–1961, 2021.

[48] K. He, X. Zhang, S. Ren, and J. Sun, ''Deep residual learning for image recognition,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 770–778.

[49] L.-C. Chen, Y. Zhu, G. Papandreou, F. Schroff, and H. Adam, ''Encoder–decoder with atrous separable convolution for semantic image segmentation,'' in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, Sep. 2018, pp. 801–818.

[50] F. Chollet, ''Xception: Deep learning with depthwise separable convolutions,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 1251–1258.

[51] X. Lu, W. Wang, J. Shen, D. J. Crandall, and L. Van Gool, ''Segmenting objects from relational visual data,'' *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 7885–7897, Nov. 2022.

[52] P.-T. de Boer, D. P. Kroese, S. Mannor, and R. Y. Rubinstein, ''A tutorial on the cross-entropy method,'' *Ann. Oper. Res.*, vol. 134, no. 1, pp. 19–67, Feb. 2005.

[53] Q. Yan, L. Xu, J. Shi, and J. Jia, ''Hierarchical saliency detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 1155–1162.

[54] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, ''Saliency detection via graph-based manifold ranking,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3166–3173.

[55] G. Li and Y. Yu, ''Visual saliency based on multiscale deep features,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 5455–5463.

[56] V. Movahedi and J. H. Elder, ''Design and perceptual validation of performance measures for salient object segmentation,'' in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2010, pp. 49–56.

[57] 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 *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 136–145.

[58] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, ''Frequency-tuned salient region detection,'' in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.

[59] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, ''Structure-measure: A new way to evaluate foreground maps,'' in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4548–4557.

NGO THIEN THU (Member, IEEE) received the B.S. degree in management information system from Vietnam National University, Ho Chi Minh City, in 2010. She is currently pursuing the combined Ph.D. degree with the Department of Computer Science and Engineering, Kyung Hee University. Her research interests include computer vision and deep learning. She was a recipient of the Best Paper Award from KSC 2019, South Korea.

MD. DELOWAR HOSSAIN received the B.Sc. and M.Sc. degrees from the Department of Information and Communication Engineering (ICE), Islamic University, Bangladesh, in 2004 and 2005, respectively, and the Ph.D. degree from the Department of Computer Science and Engineering, Kyung Hee University, South Korea. Currently, he is a Postdoctoral Researcher with the Department of Computer Science and Engineering, Kyung Hee University. He is also a Professor with the Department of Computer Science and Engineering, Hajee Mohammad Danesh Science and Technology University, Dinajpur, Bangladesh, where he was the Chairman, from 2011 to 2013. He was a Visiting Scholar with Infosys, Bengaluru, India. His current research interests include cloud/edge/fog computing, vehicular edge computing, big data, machine learning, and the Internet of Things. He was a recipient of the Best Paper Award from KSC 2018, KSC 2019, and KCC 2021, South Korea.

EUI-NAM HUH (Member, IEEE) received the B.S. degree from Busan National University, South Korea, the master's degree in computer science from The University of Texas at Austin, USA, in 1995, and the Ph.D. degree from Ohio University, USA, in 2002. Currently, he is a Professor with the Department of Computer Science and Engineering, Kyung Hee University, South Korea. His research interests include the diverse range of subjects, such as cloud computing, the Internet of Things, future internet, distributed real-time systems, mobile computing, big data, and security. He serves on the review board for the National Research Foundation of Korea. He has actively participated in community services for several organizations, including ICCSA, WPDRTS/IPDPS, APAN Sensor Network Group, ICUIMC, ICONI, APIC-IST, ICUFN, and SoICT, as different types of chairs.