Computers, Materials & Continua

Tech Science Press

DOI: 10.32604/cmc.2024.057833

ARTICLE

Salient Object Detection Based on Multi-Strategy Feature Optimization

Libo Han[1,2], Sha Tao[1,2], Wen Xia[3], Weixin Sun[3], Li Yan[3] and Wanlin Gao[1,2,3,*]

[1]College of Information and Electrical Engineering, China Agricultural University, Beijing, 100083, China

[2]Key Laboratory of Agricultural Informatization Standardization, Ministry of Agriculture and Rural Affairs, China Agricultural University, Beijing, 100083, China

[3]Pu'er University, Pu'er, 665000, China

*Corresponding Author: Wanlin Gao. Email: wanlingaocau@163.com

Received: 28 August 2024 Accepted: 14 November 2024 Published: 17 February 2025

ABSTRACT

At present, salient object detection (SOD) has achieved considerable progress. However, the methods that perform well still face the issue of inadequate detection accuracy. For example, sometimes there are problems of missed and false detections. Effectively optimizing features to capture key information and better integrating different levels of features to enhance their complementarity are two significant challenges in the domain of SOD. In response to these challenges, this study proposes a novel SOD method based on multi-strategy feature optimization. We propose the multi-size feature extraction module (MSFEM), which uses the attention mechanism, the multi-level feature fusion, and the residual block to obtain finer features. This module provides robust support for the subsequent accurate detection of the salient object. In addition, we use two rounds of feature fusion and the feedback mechanism to optimize the features obtained by the MSFEM to improve detection accuracy. The first round of feature fusion is applied to integrate the features extracted by the MSFEM to obtain more refined features. Subsequently, the feedback mechanism and the second round of feature fusion are applied to refine the features, thereby providing a stronger foundation for accurately detecting salient objects. To improve the fusion effect, we propose the feature enhancement module (FEM) and the feature optimization module (FOM). The FEM integrates the upper and lower features with the optimized features obtained by the FOM to enhance feature complementarity. The FOM uses different receptive fields, the attention mechanism, and the residual block to more effectively capture key information. Experimental results demonstrate that our method outperforms 10 state-of-the-art SOD methods.

KEYWORDS

Salient object detection; multi-strategy; feature optimization; feedback mechanism

1 Introduction

For the salient object in an image, it can quickly attract people's visual attention [1,2]. We can detect the salient object quickly and accurately using the salient object detection (SOD) method. SOD plays a significant role in many tasks, such as image compression [1], object recognition [2], object tracking [3], robot navigation [4], and image segmentation [5,6].

Currently, SOD based on natural images has been widely studied. Traditional SOD methods need to manually design feature extractors to obtain the salient object [7–9]. Although traditional SOD methods have good performance, they are difficult to extract the salient object in a complex scene accurately. In recent years, the advent of the fully convolutional network (FCN) [10] has made the FCN-based SOD methods become a hot research topic [11,12]. FCN-based SOD methods no longer need to manually design feature extractors and can automatically capture useful information. At present, many scholars have designed many methods to detect the salient object. Zhang et al. [13] designed a trifurcated cascaded refinement network, which detects the salient object by methods such as noise suppression, multi-level feature fusion, and context-aware feature learning. Gupta et al. [14] designed a gate-based context extraction module to detect the salient object. Liu et al. [15] designed feature aggregation and global guidance modules to fuse different information and guide location information. Tan et al. [16] designed a collaborative complementarity module to better capture key information. Xu et al. [17] designed a multi-source feature extraction network, which detects the salient object by methods such as feature fusion, edge detection, and saliency map refinement. Lad et al. [18] used the wavelet scattering network to capture the texture information to improve the detection accuracy. Sun et al. [19] proposed a selective feature fusion network to better fusion features. Although some existing methods show decent performance, they still have the problem of insufficient detection accuracy.

Currently, SOD faces two difficult problems. 1. The extracted features often contain a lot of useless information, which causes serious interference to the subsequent accurate detection of the salient object. How to optimize the features more effectively to better capture the key information is a difficult problem. 2. Different levels of features contain complementary information. How to better integrate these features to make them better complement is a difficult problem. Aiming at these problems, we propose a new SOD method based on multi-strategy feature optimization. In the proposed method, the multi-size feature extraction module (MSFEM) is proposed to optimize different levels of features. Low-level features reflect the original characteristics of features. High-level features contain a large amount of semantic information, they are crucial for understanding the deeper meaning of complex scenes. The residual block can be well used to capture the deeper level of semantic information. The attention mechanism can help the model to better deal with the key information. The MSFEM divides different features into 4 levels. By fusing the first 3 levels of features using the attention mechanism and summation and using residual blocks for further processing, the finer low-level features of different sizes can be obtained. By processing the 4th level of features by the residual block, the finer high-level features can be obtained. Compared to the commonly used method of directly applying different levels of features to the next modules, the MSFEM can obtain finer features to better apply to the next modules. Inspired by [20] and [21], we use two rounds of feature fusion and the feedback mechanism to obtain the final salient object by further processing the features obtained by the MSFEM. Two rounds of feature fusion and the feedback mechanism can be well used to further optimize the features to improve the detection accuracy. Although the existing two rounds of feature fusion and the feedback mechanism have excellent performance, the feature fusion methods have some drawbacks. The method of Wu et al. [20] lacks attention to the target object, which is not conducive to processing key information, and does not further optimize the fused features to capture deeper semantic information. The method of Wei et al. [21] also faces the same problem. Aiming at these problems, inspired by [20] and [21], we propose the feature enhancement module (FEM) to better realize feature fusion. The FEM first fuses the upper and lower features to achieve bi-directional enhancement of features, then fuses the features obtained by the feature optimization module (FOM) to better focus on the target object, and finally uses residual blocks to capture the

deeper level of semantic information. For the FOM, different receptive fields, the attention mechanism, and the residual block are used, which are used to focus on the target object to better deal with the key information. The FEM can better realize feature fusion compared with the existing methods.

In brief, the main contributions of this study are as follows:

(1) The MSFEM is designed. It can optimize different levels of features to better extract finer features.
(2) Inspired by [20] and [21], two rounds of feature fusion and the feedback mechanism are used. To improve the effectiveness of feature fusion, the FEM is designed. It can integrate the upper and lower features with the optimized features obtained by the FOM to better perform feature fusion.
(3) The FOM is designed by using multiple techniques such as different receptive fields, the attention mechanism, and the residual block. It can improve the model's ability to focus on key information.
(4) A new SOD method based on multi-strategy feature optimization is proposed, which can detect the salient object better than 10 state-of-the-art SOD methods.

2 Related Work

2.1 Traditional SOD Methods

Traditional SOD methods can detect the salient object through manually designed feature extractors. Wang et al. [22] designed a novel SOD method based on background driven. In this method, a saliency estimation module is proposed using the background as a cue, and the salient object is refined by using background prior information. Zhang et al. [23] designed a novel SOD method based on compactness and objectness cues. In this method, multiple strategies such as discovering the potential object, measuring compactness, and eliminating the boundary background are used to better detect the salient object. Wang et al. [24] used compactness and foreground connectivity to detect the salient object. Zhang et al. [25] designed a novel SOD method based on recursive sparse representation. In this method, the foreground and background dictionaries are used to reconstruct the error to better detect the salient object. Huang et al. [26] designed a novel SOD method based on multiple instance learning, which transforms the SOD problem into a multi-instance learning task to better detect the salient object. Nouri et al. [27] designed a novel SOD method based on multi-graph. In this method, the high contrast graph, the global graph, and the local graph are used to better detect the salient object. Srivastava et al. [28] designed a novel SOD method. In this method, gabor filters, minimum directional backgroundness, background subtraction, and objectness are used to better detect the salient object. Nouri et al. [29] designed a novel SOD method based on random graph. In this method, multiple methods such as simple linear iterative clustering and saliency labels production are used to better detect the salient object. Chen et al. [30] used the spectral graph to better extract the salient object. Naqvi et al. [31] designed a novel SOD framework based on exploiting color coefficients. In this framework, multiple methods such as region-based combination and color space selection are used to better detect the salient object. Xiao et al. [32] used eye tracking data and combined it with superpixel segmentation to better detect the salient object.

Although some existing methods have shown good performance in some situations, they are often difficult to accurately detect the salient object when faced with complex scenes.

2.2 FCN-Based SOD Methods

Unlike the traditional SOD method, the FCN-based SOD method can automatically learn useful features to detect the salient object. Currently, FCN-based SOD is also a mainstream research direction. Wu et al. [20] used multiple strategies such as self-refine, feedback, and fusion to better detect the salient object. Yang et al. [33] used feature refinement and multi-branch feature fusion to better detect the salient object. Yang et al. [34] used the feature fusion module, the contour enhancement module, and the gate module to better detect the salient object. Zhu et al. [35] designed a model that can capture context information well to better detect the salient object. Fang et al. [36] designed a ladder context correlation complementary model, which can utilize contextual information very well. Zhang et al. [37] designed a heatmap and edge guidance network. In this method, multiple modules are used to process different levels of features to better detect the salient object. Zhu et al. [38] designed a perception-and-regulation network that can capture key information very well. Huang et al. [39] designed a cross-scale resolution consistent model. In this method, multiple modules such as the cross-scale fusion module and the residual refinement module are used to better detect the salient object. Li et al. [40] designed negative and positive attention modules to better detect the salient object. Xu et al. [41] designed an adaptive fusion module based on the attention mechanism to better detect the salient object. Wu et al. [42] designed different methods to learn foreground, background, and edge features to better detect the salient object. Wu et al. [43] designed an extremely downsampled network. This network can learn the global view of the image very well. Han et al. [44] designed a residual dense collaborative network. This network has decent performance. Xu et al. [45] designed an attention-guided multiscale module and a bi-refinement module to better detect the salient object. Li et al. [46] designed a spatial frequency enhancement module. This module can better utilize spatial frequency information. Zhang et al. [47] used edge features to guide the model to better detect the salient object. Zhu et al. [48] used the method of divide-and-conquer to better perform the SOD task. Lee et al. [49] designed an adaptive graph convolution module to better deal with complex scenes. Bi et al. [50] used supervised contrastive learning to improve detection accuracy. Zhu et al. [51] used two deep prior cues to better aggregate multi-level features. Li et al. [52] refined the features from multiple perspectives to improve the accuracy. Wang et al. [53] designed a multiple enhancement network to better deal with complex scenes with cluttered backgrounds and multiple objects. Yi et al. [54] designed a two-stream gated progressive optimization network to better process the features.

Although some existing methods show decent performance, they still have the problem of insufficient detection accuracy. This is mostly due to the following two reasons. 1. The features are not sufficiently optimized, which leads to the key information can not be well captured. 2. The features are not well integrated, which leads to the essential properties of features cannot be well captured. Aiming at these problems, we propose a new SOD method based on multi-strategy feature optimization, which has excellent performance.

3 Proposed Method

Fig. 1 shows the overall framework of the proposed method. At present, many SOD methods [20,21,33,35] are realized based on ResNet-50 [55], and have achieved good effects. Therefore, five levels of features are first extracted using ResNet-50, which are denoted as r_1, r_2, r_3, r_4, and r_5, respectively. Since r_1 contains more noise and affects the computational efficiency due to the larger scale, r_2, r_3, r_4, and r_5 are used to obtain the salient object. To facilitate the subsequent processing, the 1×1 convolution, batch normalization, and relu operation are implemented sequentially to change the number of channels of these features to 192, thus obtaining φ_1, φ_2, φ_3, and φ_4. Next, the MSFEM

is used to process φ_1, φ_2, φ_3, and φ_4 to capture finer features. Then, two rounds of feature fusion and the feedback mechanism are used to further optimize the features to improve the detection accuracy. Next, the MSFEM, two rounds of feature fusion, and the feedback mechanism are described in detail.

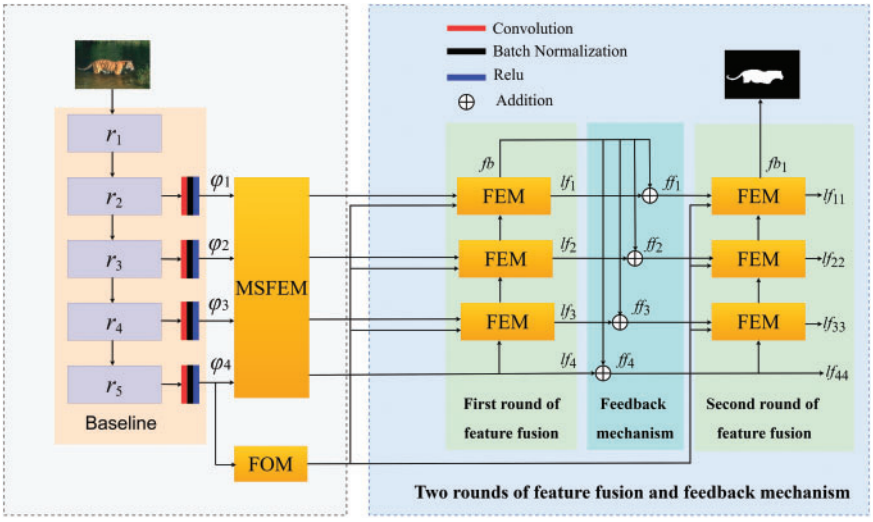

Figure 1: Overall framework of the proposed method

3.1 MSFEM

The MSFEM can refine the features well. The structure of the MSFEM is shown in Fig. 2. Next, the structure of the MSFEM is described in combination with Fig. 2.

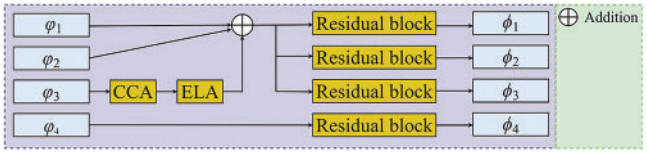

Figure 2: Structure of the MSFEM

Let φ_1, φ_2, φ_3, and φ_4 be the four levels of features of the input. Fusing different levels of low-level features can help model to capture more context information. Compared with φ_4, φ_1, φ_2, and φ_3 contain more low-level information. Therefore, φ_1, φ_2, and φ_3 are fused together. The attention model can help the model to focus on the target object to better deal with the key information. For some images, the salient object occupies a small region while the non-salient object occupies a large region. When using the attention model, if the features contain too much low-level information about the non-salient object, these information can easily have a bad effect on capturing low-level information of the salient object. Compared with φ_1 and φ_2, φ_3 has a higher level and contains less low-level information of non-salient regions. Therefore, we use the attention model to process φ_3 and fuse the obtained features with φ_1 and φ_2 to obtain η. For the attention model, the criss-cross attention model (CCA) [56] is first used to focus on the target object through contextual information, and then more attention is given to the target object by focusing on the local region using the efficient local attention model (ELA) [57]. Let η be the fused features. The method of obtaining η is as follows:

$$\eta = \varphi_1 + TRS_{\varphi_1}(\varphi_2) + TRS_{\varphi_1}(\varsigma) \tag{1}$$

$$\varsigma = ELA\left(CCA\left(\varphi_3\right)\right) \tag{2}$$

where $TRS_{\varphi_1}(\cdot)$ denotes that the size of the image is transformed to the same size as φ_1 using bilinear interpolation, $CCA(\cdot)$ denotes that the CCA, $ELA(\cdot)$ denotes that the ELA. Note that for the ELA, the well-performing ELA-Base is used in this study.

Since there are some differences in the size of salient objects in different images, η is changed to different sizes to obtain φ_1', φ_2', and φ_3' to better capture the information of salient objects with different sizes in subsequent operations.

$$\varphi_1' = \eta \tag{3}$$

$$\varphi_2' = TRS_{\varphi_2}\left(\eta\right) \tag{4}$$

$$\varphi_3' = TRS_{\varphi_3}\left(\eta\right) \tag{5}$$

where $TRS_{\varphi_2}(\cdot)$ denotes that the size of the image is transformed to the same size as φ_2 using bilinear interpolation, and $TRS_{\varphi_3}(\cdot)$ denotes that the size of the image is transformed to the same size as φ_3 using bilinear interpolation.

The residual block can be well used to capture the deeper level of semantic information and can help to improve the model's performance [55]. Therefore, we use residual blocks to process φ_1', φ_2', φ_3', and φ_4 to capture the deeper level of semantic information. Let ϕ_1, ϕ_2, ϕ_3, and ϕ_4 be the four levels of features of the output. ϕ_1, ϕ_2, ϕ_3, and ϕ_4 are obtained in the following methods:

$$\phi_1 = RES\left(\varphi_1'\right) \tag{6}$$

$$\phi_2 = RES\left(\varphi_2'\right) \tag{7}$$

$$\phi_3 = RES\left(\varphi_3'\right) \tag{8}$$

$$\phi_4 = RES\left(\varphi_4\right) \tag{9}$$

where $RES(\cdot)$ denotes the residual block. Inspired by [55], the residual block shown in Fig. 3 is used in this study. Sometimes the captured salient object is larger than the ground truth. Therefore, a 1-pixel expansion of the feature's boundaries is done in all the 1×1 convolutions. By subsequent scaling, the captured salient object becomes smaller accordingly. By fusing the obtained features with the original features can better capture deeper semantic information. Let ρ be the input features, and ρ' be the output features. The residual block is realized as follows:

$$\rho' = RL\left(\rho + TRS_{\rho}\left(C_1\left(CRL_3\left(CRL_1\left(\rho\right)\right)\right)\right)\right) \tag{10}$$

where $CRL_1(\cdot)$ denotes that the 1×1 convolution and relu operation are implemented sequentially, $CRL_3(\cdot)$ denotes that the 3×3 convolution and relu operation are implemented sequentially, $C_1(\cdot)$ denotes that the 1×1 convolution is performed, $TRS_{\rho}(\cdot)$ denotes that the size of the image is transformed to the same size as ρ using bilinear interpolation, and $RL(\cdot)$ denotes the relu operation.

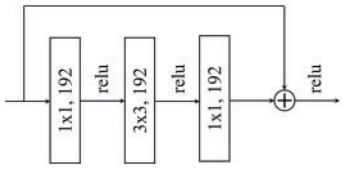

Figure 3: Residual block

3.2 Two Rounds of Feature Fusion and the Feedback Mechanism

Two rounds of feature fusion and the feedback mechanism can improve the detection accuracy by optimizing the features obtained by the MSFEM. The advantages of two rounds of feature fusion and the feedback mechanism are as follows [20,21]. The first round of feature fusion can obtain refiner features well, but these features often contain some useless information. The better results can be obtained by feedback these features into different levels of features to do the second fusion. Due to the feature fusion methods used by existing methods lack attention to the target object, which is not conducive to processing key information. In addition, these fusion methods do not further optimize the fused features to capture deeper semantic information. Therefore, this study proposes a new feature fusion method applied to two rounds of feature fusion and combines them with the feedback mechanism to detect the salient object. As shown in Fig. 1, two rounds of feature fusion and the feedback mechanism used in this study are mainly composed of the FEM, the FOM, and the feedback mechanism. The FOM uses different receptive fields, the attention mechanism, and the residual block to better capture the key information. The FEM first fuses the upper and lower features to achieve bi-directional enhancement of features, then fuses the features obtained by FOM to better focus on the target object, and finally uses residual blocks to capture the deeper level of semantic information. The feedback mechanism is used to feed the finer features obtained from the first round of feature fusion to the second round of feature fusion for further optimization of the features. The salient object can be well captured by the above modules working in collaboration with each other. Next, the FEM, the FOM, and the feedback mechanism are described in detail.

3.2.1 FEM

The FFM can improve the detection accuracy by enhancing the features at different levels. Fig. 4 shows the structure of the FEM. The FEM realizes feature fusion between different levels of features by multiplication and fuses the fused features with the original different levels of features again to realize bidirectional enhancement of features. Furthermore, *opf* obtained by the FOM in Section 3.2.2 is added to the fused features to better focus on key information. Finally, the residual block described in Section 3.1 is used to capture the deeper level of semantic information. Let τ_1 be the features with a higher level in the neighboring levels of features, τ_2 be the features with a lower level in the neighboring levels of features, χ_1 be the features with a higher level of the output, and χ_2 be the features with a lower level of the output. The FEM is realized as follows:

$$\chi_2 = RES\left(CBRL_3\left(CBRL_3\left(fuse\right) + \tau_2^{'} + opff\right)\right) \tag{11}$$

$$\chi_1 = RES\left(CBRL_3\left(CBRL_3\left(fuse\right) + \tau_1^{'} + opff\right)\right) \tag{12}$$

$$fuse = CBRL_3\left(\tau_2^{'}\right) \times CBRL_3\left(\tau_1^{'}\right) \tag{13}$$

$$opff = CBRL_3\left(TRS_{\tau_2}\left(opf\right)\right) \tag{14}$$

$$\tau_1^{'} = CBRL_3\left(TRS_{\tau_2}\left(\tau_1\right)\right) \tag{15}$$

$$\tau_2^{'} = CBRL_3\left(\tau_2\right) \tag{16}$$

where $CBRL_3(\cdot)$ denotes that the 3×3 convolution, batch normalization, and relu operation are implemented sequentially, $TRS_{\tau_2}(\cdot)$ denotes that the size of the image is transformed to the same size as τ_2 using bilinear interpolation, and $RES(\cdot)$ denotes the residual block.

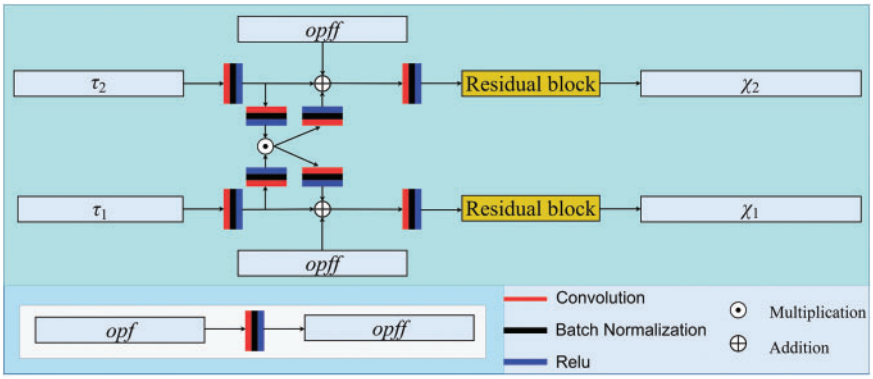

Figure 4: Structure of the FEM

3.2.2 FOM

The FOM can improve the detection accuracy by effectively capturing key information. Fig. 5 shows the structure of the FOM. For the FOM, the atrous spatial pyramid pooling (ASPP) module [58] is first used to capture richer contextual information through different receptive fields. Then, the attention models are used to focus on the target object. The CCA is first used to focus on the target object through contextual information, and then more attention is given to the target object by focusing on the local region using the ELA. Finally, the residual block described in Section 3.1 is used to obtain *opf* to capture the deeper level of semantic information. Let φ_4 be the input features.

$$opf = RES\left(ELA\left(CCA\left(CBRL_1\left(ASPP\left(CBRL_1\left(\varphi_4\right)\right)\right)\right)\right)\right) \tag{17}$$

where $CBRL_1(\cdot)$ denotes that the 1×1 convolution, batch normalization, and relu operation are implemented sequentially, $ASPP(\cdot)$ denotes the ASPP module, and $RES(\cdot)$, $CCA(\cdot)$, and $ELA(\cdot)$ denote the residual block, the CCA, and the ELA, respectively. Before using the ASPP module, $CBRL_1(\cdot)$ is used to convert the number of feature channels from 192 to 256, and after using the ASPP module, $CBRL_1(\cdot)$ is used to convert the number of feature channels from 256 to 192.

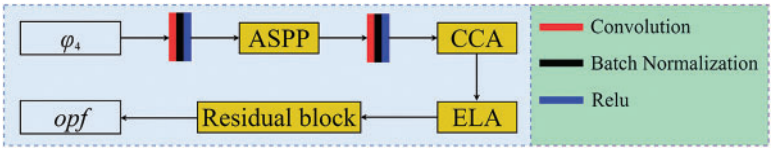

Figure 5: Structure of the FOM

3.2.3 Feedback Mechanism

After completing the first round of feature fusion, the finer features can be well obtained. These features contain rich information such as edges, semantics, etc. The above features are feedback to all levels of features for the second round of feature fusion can well improve the detection accuracy. In this study, the existing feedback mechanism is uesd. Next, the feedback mechanism is described in combination with Fig. 1. As shown in Fig. 1, let *fb* be the feedback features, lf_j be the fusion features with a lower level at the *j*th level, and ff_j be the *j*th level of features obtained by fusing lf_j and *fb*. The feedback mechanism is realized as follows:

$$ff_j = lf_j + TRS_{lf_j}\left(fb\right) \tag{18}$$

where $j = 1, 2, 3, 4$, and $TRS_{lf_j}(\cdot)$ denotes that the size of the image is transformed to the same size as lf_j using bilinear interpolation.

3.3 Loss Function

In this study, the loss function of Wei et al. [21] is used. Compared with the commonly used binary cross entropy (BCE) and intersection over union (IOU), the weighted BCE and the weighted IOU used by Wei et al. have better performance. Next, the loss function used is described in combination with Fig. 1. As shown in Fig. 1, let fb and fb_1 be the fusion features with a higher level at the 2th level, and lf_{11}, lf_{22}, lf_{33}, and lf_{44} be the features with a lower level at the 2th, 3th, 4th, and 5th levels, respectively. The total loss LF is calculated by Eq. (19).

$$LF = LF_1 + LF_2 \tag{19}$$

$$LF_1 = \frac{1}{K}\sum_{k=1}^{K} LFAD_k \tag{20}$$

$$LF_2 = \sum_{t=2}^{5} \frac{1}{2^{t-1}} LFAD_t \tag{21}$$

$$LFAD = LF_{wbce} + LF_{wiou} \tag{22}$$

where $K = 2$, LF_{wbce} denotes the weighted BCE, LF_{wiou} denotes the weighted IOU, LF_1 calculates the loss through fb and fb_1, LF_2 calculates the loss through lf_{11}, lf_{22}, lf_{33}, and lf_{44}.

4 Experimental Results

4.1 Datasets and Evaluation Metrics

We have conducted extensive experiments on four mainstream databases to verify the superiority of our designed model. These four databases are the ECSSD [59], DUTS [60], DUT-OMRON [61], and HKU-IS [62] databases, respectively. The ECSSD database contains 1000 images. The DUTS database includes the training and test sets, which are the DUTS-TR and DUTS-TE databases, respectively. The DUTS-TR database contains 10,553 images. The DUTS-TE database contains 5019 images. The DUT-OMRON database contains 5168 images. The HKU-IS database contains 4447 images. The images in these databases all contain a large number of complex scenes. These complex scenes greatly increase the difficulty in detecting the salient object.

We used seven evaluation metrics to compare different methods objectively. These metrics are often applied to the evaluation of SOD methods. They are mean absolute error (MAE) [63], E-measure (EM) [64], S-measure (SM) [65], weighted F-measure (WFM) [66], precision-recall curves, parameter quantity, and frames per second (FPS), respectively. MAE can be used to evaluate the difference between the ground truth G and the saliency map S.

$$MAE = \frac{1}{I \times J}\sum_{i=1}^{I}\sum_{j=1}^{J} |S(i,j) - G(i,j)| \tag{23}$$

where I and J represent the length and width of G, respectively. EM is a good evaluation metric and has been used by many scholars [36–38]. When using EM, the adaptive threshold [67] can be used to obtain the binary foreground map. The adaptive threshold may be greater than 1. In this study, we set the adaptive threshold greater than 1 to 1. In addition, we use the value greater than or equal to the

adaptive threshold to obtain the foreground. SM can be used to evaluate the structural similarity. Let S_r and S_{os} be the region-aware structural similarity and the object-aware structural similarity, respectively. SM is calculated as follows:

$$S_m = (1 - \kappa) \times S_r + \kappa \times S_{os} \tag{24}$$

where κ is set 0.5. WFM is a good evaluation metric and has been used by many scholars [33,38,47]. Precision-recall curves [20,21] can be used to display the overall performance. Parameter quantity and FPS can be used to evaluate the complexity of the model.

4.2 Implementation Details

In this study, we used the DUTS-TR database to train our model. Pytorch was used to implement our method. The model was trained by RTX 4090 GPU and tested by GTX 1080 Ti GPU. When training the whole network, the batch size, the weight decay, the momentum, and the maximum epoch were set to 32, 5e−4, 0.9, and 48, respectively. Furthermore, following the method of Wei et al. [21], multi-scale input, horizontal flip, and random crop were used for data augmentation, stochastic gradient descent was used train the whole network, the backbone network used the pre-trained ResNet-50 [55] model, the learning rate was adjusted by warm-up and linear decay strategies, ResNet-50 backbone and other parts of the maximum learning rate were set to 0.005 and 0.05, respectively, and during testing, the size of each image was modified to 352×352. In addition, the random seed was set to 42.

4.3 Ablation Study

To better understand the impact of each module in our method on the overall performance, we designed different schemes for validation. MAE, EM, SM, and WFM were used for objective evaluation. Meanwhile, we also performed visual comparisons. In addition, we used the DUTS-TR database as the training set to obtain the models of different schemes and used the DUTS-TE and DUT-OMRON databases to test these models.

Table 1 shows the objective evaluation results obtained by different schemes. A denotes that only the baseline model labeled in Fig. 1 is used. The detection result is output after φ_4. B denotes that the baseline model and the FEM with FOM removed are used. C denotes that the baseline model and the FEM are used. D denotes that the baseline model, the FEM, and the MSFEM are used. Compared with A, B can provide better results. This shows that the performance of the model can be well improved when combining the baseline model with the FEM with the FOM removed. Compared with B, C can provide better results. This shows that adding the FOM to the FEM can better improve the performance of the model. Compared with C, D can provide better results. This shows that the MSFEM can improve the performance of the model very well. Fig. 6 shows visual effects obtained by different schemes. The results of the visual assessment are similar to those of the objective assessment. Compared with A, B can provide better results. Compared with B, C can provide better results. Compared with C, D can provide better results. Based on the above analysis, we can know that the FOM, the FEM, and the MSFEM can improve the performance of the model, and it is necessary to combine these modules.

Table 1: Ablation studies on the DUTS-TE and DUT-OMRON datasets

Schemes	Baseline	FEM(-FOM)	FEM	MSFEM	DUTS-TE				DUT-OMRON			
					MAE↓	EM↑	SM↑	WFM↑	MAE↓	EM↑	SM↑	WFM↑
A	√				0.0584	0.8630	0.8336	0.7248	0.0776	0.8243	0.7912	0.6558
B	√	√			0.0359	0.9221	0.8919	0.8414	0.0556	0.8746	0.8395	0.7511
C	√		√		0.0331	0.9270	0.8948	0.8489	0.0520	0.8748	0.8409	0.7549
D	√		√	√	0.0332	0.9285	0.8966	0.8507	0.0518	0.8792	0.8440	0.7600

Note: The symbol ↓ means the evaluation value is smaller, the model's performance is better. The symbol ↑ means the evaluation value is larger, the model's performance is better. The values ranked first, second, and third are highlighted by red color, blue color, and green color, respectively.

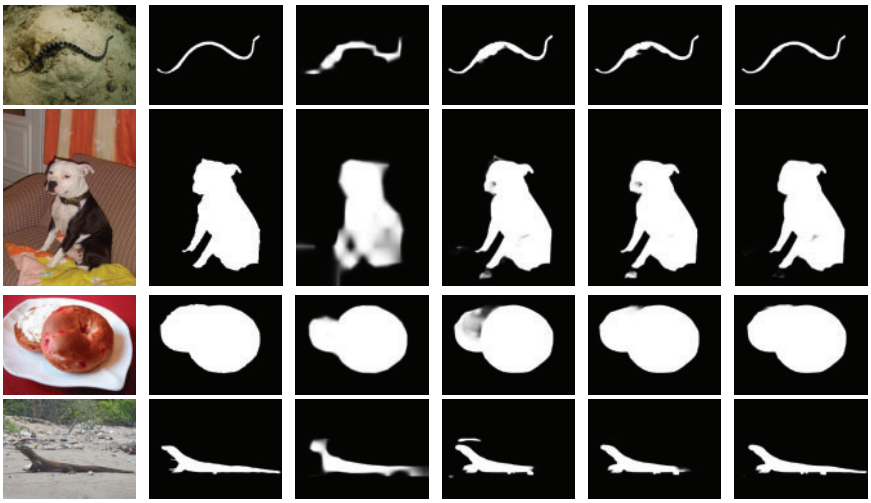

Figure 6: Visual effects obtained by different schemes

4.4 Objective Comparison

In this study, SRFFNet [20], F^{3}Net [21], PurNet [68], CSNet [69], GFINet [70], SCRN [71], CPD [72], DCENet [73], ITSD [74], and BANet [75] were used for objective comparison. To make a fair comparison, we used the saliency maps provided by the authors or the code published by the authors to obtain the saliency maps.

Fig. 7 shows the precision-recall curves obtained on different databases. As shown in Fig. 7, the results obtained by the proposed method on each database are superior to most of the methods. Overall, the performance of the proposed method is excellent. Table 2 shows the results of the objective comparison obtained by the other four metrics. For the ECSSD database, except for GFINet and SRFFNet, the number of metrics ranked second obtained by our method, PurNet, and SCRN is the most. For the DUTS-TE database, the number of metrics ranked first obtained by our method is far more than other methods. For the DUTS-OMRON database, the total number of metrics ranked first and second obtained by our method is far more than other methods. For the HKU-IS database, the

number of metrics ranked first obtained by our method is far more than other methods. In conclusion, our method has a better performance.

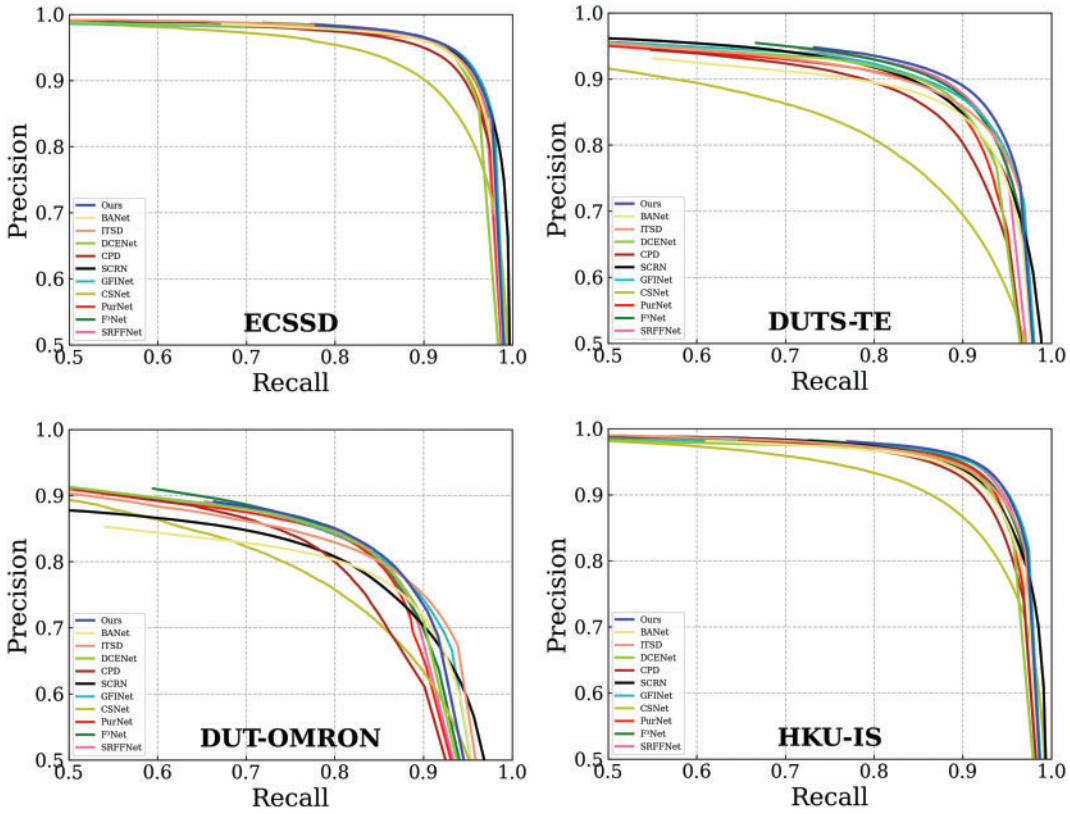

Figure 7: Precision-recall curves

Table 2: Objective comparison with 10 advanced methods

Methods	ECSSD				DUTS-TE			
	MAE↓	EM↑	SM↑	WFM↑	MAE↓	EM↑	SM↑	WFM↑
SRFFNet	0.0324	0.9486	0.9233	0.9160	0.0345	0.9300	0.8884	0.8491
F^3Net	0.0333	0.9461	0.9242	0.9122	0.0354	0.9183	0.8884	0.8349
PurNet	0.0346	0.9528	0.9245	0.9070	0.0389	0.9147	0.8814	0.8170
CSNet	0.0645	0.8988	0.8928	0.8062	0.0744	0.8226	0.8225	0.6448
GFINet	0.0316	0.9535	0.9288	0.9156	0.0376	0.9166	0.8892	0.8320
SCRN	0.0375	0.9424	0.9272	0.8995	0.0397	0.9015	0.8849	0.8031
CPD	0.0370	0.9494	0.9181	0.8980	0.0433	0.9037	0.8691	0.7955
DCENet	0.0347	0.9515	0.9214	0.9126	0.0379	0.9216	0.8823	0.8337
ITSD	0.0345	0.9316	0.9248	0.9105	0.0410	0.8979	0.8849	0.8236
BANet	0.0349	0.9527	0.9239	0.9078	0.0398	0.9070	0.8787	0.8109
Ours	0.0321	0.9477	0.9257	0.9155	0.0332	0.9285	0.8966	0.8507

(Continued)

Table 2 (continued)

Methods	DUT-OMRON				HKU-IS			
	MAE↓	EM↑	SM↑	WFM↑	MAE↓	EM↑	SM↑	WFM↑
SRFFNet	0.0525	0.8812	0.8386	0.7588	0.0271	0.9607	0.9199	0.9107
F^3Net	0.0526	0.8763	0.8385	0.7473	0.0280	0.9581	0.9172	0.9004
PurNet	0.0512	0.8762	0.8413	0.7473	0.0298	0.9559	0.9177	0.8919
CSNet	0.0807	0.8163	0.8050	0.6206	0.0586	0.9189	0.8815	0.7773
GFINet	0.0539	0.8769	0.8437	0.7560	0.0269	0.9604	0.9230	0.9065
SCRN	0.0560	0.8688	0.8365	0.7202	0.0337	0.9528	0.9158	0.8758
CPD	0.0557	0.8722	0.8247	0.7192	0.0331	0.9516	0.9086	0.8791
DCENet	0.0550	0.8735	0.8386	0.7535	0.0293	0.9573	0.9154	0.8976
ITSD	0.0608	0.8669	0.8403	0.7499	0.0307	0.9529	0.9170	0.8939
BANet	0.0587	0.8654	0.8322	0.7356	0.0323	0.9544	0.9131	0.8863
Ours	0.0518	0.8792	0.8440	0.7600	0.0263	0.9619	0.9234	0.9104

Note: The symbol ↓ means the evaluation value is smaller, the model's performance is better. The symbol ↑ means the evaluation value is larger, the model's performance is better. The values ranked first and second are highlighted by blue color and green color, respectively.

To further validate the performance of the model, we evaluated the complexity of the proposed model through FPS and the model's parameter quantity. In addition, a comparison was performed with SRFFNet. The reasons for choosing SRFFNet are as follows. As shown in Table 2, except for the ECSSD database, the number of metrics ranked first obtained by SRFFNet is second. In addition, from Table 2, we can find that SRFFNet is ranked second in terms of performance when all databases are considered. Note that since both the proposed method and SRFFNet used the pre-trained ResNet-50 model, we counted the total parameters of the other modules. In addition, we obtained FPS values according to the DUTS-TE and DUT-OMRON databases. As shown in Table 3, although the results obtained by the proposed method are worse than those of SRFFNet, it still has a high FPS value. The FPS value obtained by the proposed method on GTX 1080 Ti GPU is 19 fps, indicating that 19 images can be processed per second. This indicates that the proposed method can be applied to real-time detection and large-scale datasets. Combined with Fig. 7 and Tables 2 and 3, it can be seen that the proposed method can detect the salient objects more accurately while having a high efficiency.

Table 3: Model complexity comparison

	Parameter quantity (M)	FPS (fps)
SRFFNet	27.56	23
Ours	51.54	19

4.5 Visual Effect and Objective Analyze

In this part, we validate the performance of our method by presenting visual effects and in-depth objective analysis.

Fig. 8 shows the visual effects of the detected salient objects in different scenes. In this study, Img1–Img11 are used to denote the original images from the first row to the 11th row. For Img1 and Img8,

each image contains a large number of complex backgrounds, and the salient object to be detected in each image is also very small. Our method can detect these salient objects very well. For Img6, it contains a large number of complex backgrounds, and the salient object to be detected is large. Our method can also detect this salient object very well. Of course, there are many other sizes of images, such as Img2, Img7, and Img11. Our method can still detect these salient objects very well. In addition, for Img10, the colors of salient objects are very similar to the colors of some backgrounds. Our method can still detect these salient objects well. The objective evaluation values are presented in Fig. 9. In this study, 1–11 are used to denote Img1–Img11. As shown in Fig. 9, our method can better obtain smaller MAE values and larger EM, SM, and WFM values. Our method can obtain good objective evaluation results.

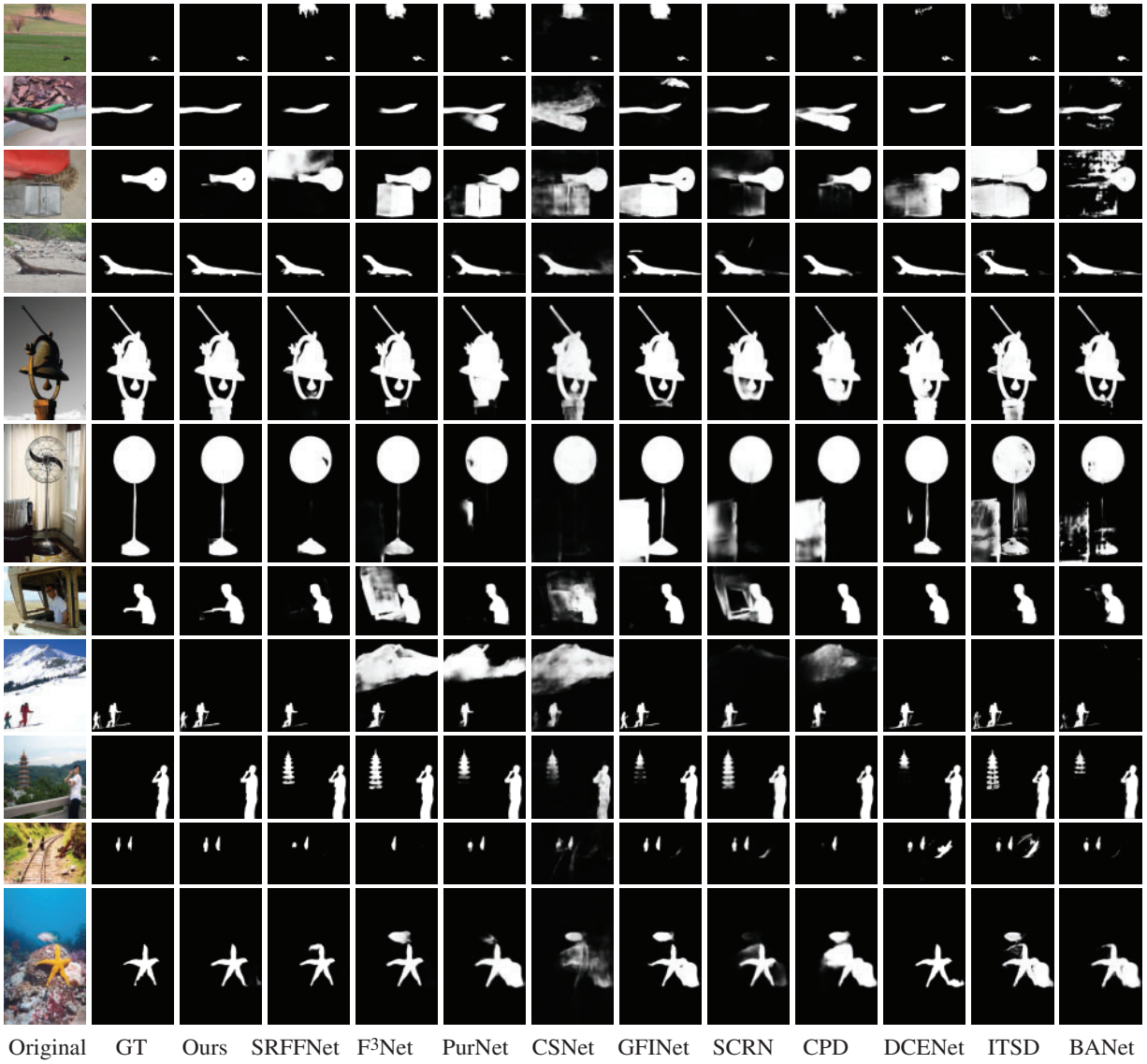

Figure 8: Visual effects of our method and 10 state-of-the-art methods. GT denotes the ground truth

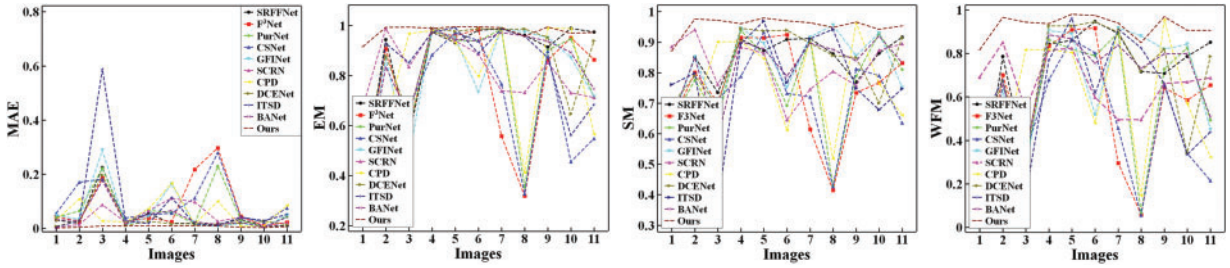

Figure 9: Objective comparison with 10 state-of-the-art methods

5 Conclusion and Discussion

To detect the salient object more accurately, this study proposes a new SOD method based on multi-strategy feature optimization. The MSFEM is proposed to refine the features. Meanwhile, the MSFEM, two rounds of feature fusion, and the feedback mechanism are well combined to realize the enhancement of feature expression capability and the deep fusion of multi-level features. To improve the performance of two rounds of feature fusion, the FEM and the FOM are proposed. The FEM can well achieve bi-directional enhancement of features and focus on key information by combining the features obtained by the FOM. Meanwhile, the FEM uses residual blocks to learn deeper semantic information. Numerous experimental results show that our method can perform better than 10 state-of-the-art SOD methods.

In addition, numerous experiments show that our method has good generalization ability to natural images. The databases used in this study are mainstream databases, which are obtained according to practical needs with the development of SOD. These databases contain a large number of different types of complex scenes. The proposed method has good detection ability in all four mainstream databases, which means that it has good generalization ability for dealing with different types of natural images. Of course, the proposed method is proposed based on natural images. It may not have good detection ability in dealing with unnatural images, such as remote sensing images. Next, we will propose specific models for specific unnatural images. Moreover, although the proposed method has high detection accuracy and high efficiency, it has higher complexity compared to some state-of-the-art methods. Next, we will investigate how to improve the detection accuracy with low complexity.

Acknowledgement: Thanks to all those who helped us complete this study.

Funding Statement: The authors received no specific funding for this study.

Author Contributions: The authors confirm contribution to the paper as follows: Libo Han is mainly responsible for the writing of the paper and the program design. Sha Tao is mainly responsible for the revision and guidance of the paper. Wen Xia, Weixin Sun, and Li Yan provide support for the project. Wanlin Gao is mainly responsible for the project administration and the revision and guidance of the paper. All authors reviewed the results and approved the final version of the manuscript.

Availability of Data and Materials: The data and materials of this study are available from the corresponding author upon reasonable request.

Ethics Approval: Not applicable.

Conflicts of Interest: The authors declare no conflicts of interest to report regarding the present study.

References

[1] C. X. Xia, Y. G. Sun, X. J. Gao, B. Ge, and S. S. Duan, "DMINet: Dense multi-scale inference network for salient object detection," *Vis. Comput.*, vol. 38, no. 9–10, pp. 3059–3072, 2022. doi: 10.1007/s00371-022-02561-8.

[2] Q. Zhang, M. X. Duanmu, Y. J. Luo, Y. Liu, and J. G. Han, "Engaging part-whole hierarchies and contrast cues for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 6, pp. 3644–3658, 2022. doi: 10.1109/TCSVT.2021.3104932.

[3] I. Ullah *et al.*, "CMGNet: Context-aware middle-layer guidance network for salient object detection," *J. King Saud Univ.-Comput. Inf. Sci.*, vol. 36, no. 1, 2024, Art. no. 101838. doi: 10.1016/j.jksuci.2023.101838.

[4] H. Huang *et al.*, "Multi-feature aggregation network for salient object detection," *Signal Image Video Process.*, vol. 17, no. 4, pp. 1043–1051, 2023. doi: 10.1007/s11760-022-02310-3.

[5] W. Fang, Y. X. Fu, and V. S. Sheng, "Dual backbone interaction network for burned area segmentation in optical remote sensing images," *IEEE Geosci. Remote Sens. Lett.*, vol. 21, 2024, Art. no. 6008805. doi: 10.1109/LGRS.2024.3369619.

[6] W. Fang, Y. X. Fu, and V. S. Sheng, "FPS-U2Net: Combining U2Net and multi-level aggregation architecture for fire point segmentation in remote sensing images," *Comput. Geosci.*, vol. 189, 2024, Art. no. 105628.

[7] Y. Pang, X. S. Yu, Y. Wang, and C. D. Wu, "Salient object detection based on novel graph model," *J. Vis. Commun. Image Represent.*, vol. 65, 2019, Art. no. 102676.

[8] Y. Z. Wang and G. H. Peng, "Salient object detection via incorporating multiple manifold ranking," *Signal Image Video Process.*, vol. 13, no. 8, pp. 1603–1610, 2019.

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

[10] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 3431–3440.

[11] Y. M. Fang, H. Y. Zhang, J. B. Yan, W. H. Jiang, and Y. Liu, "UDNet: Uncertainty-aware deep network for salient object detection," *Pattern Recognit.*, vol. 134, 2023, Art. no. 109099.

[12] Y. S. Li, J. Wu, L. Zhu, and W. W. Wang, "Salient object detection based on adaptive deep differential pyramid," in *Proc. CCDC*, Yichang, China, 2023, pp. 4081–4086.

[13] Q. Zhang, R. Zhao, and L. Q. Zhang, "TCRNet: A trifurcated cascaded refinement network for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 33, no. 1, pp. 298–311, 2023. doi: 10.1109/TCSVT.2022.3199780.

[14] A. K. Gupta, A. Seal, P. Khanna, A. Yazidi, and O. Krejcar, "Gated contextual features for salient object detection," *IEEE Trans. Instrum. Meas.*, vol. 70, 2021, Art. no. 5007613. doi: 10.1109/TIM.2021.3064423.

[15] J. J. Liu, Q. B. Hou, Z. A. Liu, and M. M. Cheng, "PoolNet+: Exploring the potential of pooling for salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 887–904, 2023. doi: 10.1109/TPAMI.2021.3140168.

[16] Z. S. Tan and X. D. Gu, "Bridging feature complementarity gap between encoder and decoder for salient object detection," *Digit. Signal Process.*, vol. 133, 2023, Art. no. 103841. doi: 10.1016/j.dsp.2022.103841.

[17] K. Xu and J. C. Guo, "A multi-source feature extraction network for salient object detection," *Neural Comput. Applic.*, vol. 35, no. 35, pp. 24727–24742, 2023. doi: 10.1007/s00521-022-08172-7.

[18] B. V. Lad, M. F. Hashmi, and A. G. Keskar, "LDWS-net: A learnable deep wavelet scattering network for RGB salient object detection," *Image Vis. Comput.*, vol. 137, 2023, Art. no. 104748. doi: 10.1016/j.imavis.2023.104748.

[19] F. M. Sun, X. Yuan, and C. X. Zhao, "Selective feature fusion network for salient object detection," *IET Comput. Vis.*, vol. 17, no. 4, pp. 483–495, 2023. doi: 10.1049/cvi2.12183.

[20] S. Wu and G. J. Zhang, "SRFFNet: Self-refine, fusion and feedback for salient object detection," *Cognit. Comput.*, vol. 15, no. 3, pp. 943–955, 2023. doi: 10.1007/s12559-023-10130-x.

[21] J. Wei, S. H. Wang, and Q. M. Huang, "F^3Net: Fusion, feedback and focus for salient object detection," in *Proc. AAAI Conf. on Artificial Intelligence*, New York, NY, USA, 2020, vol. 34, pp. 12321–12328.
[22] Z. L. Wang, D. Xiang, S. H. Hou, and F. Wu, "Background-driven salient object detection," *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 750–762, 2017. doi: 10.1109/TMM.2016.2636739.
[23] Q. Zhang, J. J. Lin, W. J. Li, Y. J. Shi, and G. G. Cao, "Salient object detection via compactness and objectness cues," *Vis. Comput.*, vol. 34, no. 4, pp. 473–489, 2018. doi: 10.1007/s00371-017-1354-0.
[24] Y. Z. Wang and G. H. Peng, "Salient object detection based on compactness and foreground connectivity," *Mach. Vision Appl.*, vol. 29, no. 7, pp. 1143–1155, 2018. doi: 10.1007/s00138-018-0958-3.
[25] Y. J. Zhang, X. Wang, X. W. Xie, and Y. S. Li, "Salient object detection via recursive sparse representation," *Remote Sens.*, vol. 10, no. 4, 2018, Art. no. 652. doi: 10.3390/rs10040652.
[26] F. Huang, J. Q. Qi, H. C. Lu, L. H. Zhang, and X. Ruan, "Salient object detection via multiple instance learning," *IEEE Trans. Image Process.*, vol. 26, no. 4, pp. 1911–1922, 2017. doi: 10.1109/TIP.2017.2669878.
[27] F. Nouri, K. Kazemi, and H. Danyali, "Salient object detection using local, global and high contrast graphs," *Signal Image Video Process.*, vol. 12, no. 4, pp. 659–667, 2018. doi: 10.1007/s11760-017-1205-5.
[28] G. Srivastava and R. Srivastava, "Salient object detection using background subtraction, gabor filters, objectness and minimum directional backgroundness," *J. Vis. Commun. Image Represent.*, vol. 62, pp. 330–339, 2019. doi: 10.1016/j.jvcir.2019.06.005.
[29] F. Nouri, K. Kazemi, and H. Danyali, "Salient object detection method using random graph," *Multimed. Tools Appl.*, vol. 77, no. 19, pp. 24681–24699, 2018. doi: 10.1007/s11042-018-5668-3.
[30] J. Z. Chen *et al.*, "Salient object detection via spectral graph weighted low rank matrix recovery," *J. Vis. Commun. Image Represent.*, vol. 50, pp. 270–279, 2018. doi: 10.1016/j.jvcir.2017.12.006.
[31] S. S. Naqvi, J. Mirza, and T. Bashir, "A unified framework for exploiting color coefficients for salient object detection," *Neurocomputing*, vol. 312, pp. 187–200, 2018. doi: 10.1016/j.neucom.2018.05.091.
[32] F. Xiao, L. C. Peng, L. Fu, and X. P. Gao, "Salient object detection based on eye tracking data," *Signal Process.*, vol. 144, pp. 392–397, 2018. doi: 10.1016/j.sigpro.2017.10.019.
[33] J. Y. Yang, Y. J. Shi, J. Zhang, Q. Q. Guo, Q. Zhang, and L. Cui, "Multi-branch feature fusion and refinement network for salient object detection," *Multimed. Syst.*, vol. 30, no. 4, 2024, Art. no. 190. doi: 10.1007/s00530-024-01356-2.
[34] A. P. Yang, Y. Liu, S. M. Cheng, J. L. Cao, Z. Ji and Y. W. Pang, "Spatial attention-guided deformable fusion network for salient object detection," *Multimed. Syst.*, vol. 29, no. 5, pp. 2563–2573, 2023.
[35] G. Zhu, L. Wang, and J. P. Tang, "Learning discriminative context for salient object detection," *Eng. Appl. Artif. Intell.*, vol. 131, 2024, Art. no. 107820.
[36] X. Fang, J. C. Zhu, X. L. Shao, and H. P. Wang, "LC3Net: Ladder context correlation complementary network for salient object detection," *Knowl.-Based Syst.*, vol. 242, 2022, Art. no. 108372.
[37] B. T. Zhang, L. H. Tian, C. Li, and Y. Yang, "Heatmap and edge guidance network for salient object detection," *Computers & Electrical Engineering*, vol. 105, 2023, Art. no. 108525.
[38] J. C. Zhu, X. Y. Zhang, X. Fang, Y. X. Wang, P. L. Tan and J. N. Liu, "Perception-and-regulation network for salient object detection," *IEEE Trans. Multimedia*, vol. 25, pp. 6525–6537, 2023.
[39] X. Y. Huang, W. Liu, M. H. Li, and H. Y. Nie, "Cross-scale resolution consistent network for salient object detection," *IET Image Process.*, vol. 18, no. 10, pp. 2788–2799, 2024. doi: 10.1049/ipr2.13136.
[40] J. X. Li, Z. F. Pan, Q. S. Liu, Y. Cui, and Y. B. Sun, "Complementarity-aware attention network for salient object detection," *IEEE Trans. Cybern.*, vol. 52, no. 2, pp. 873–886, 2022. doi: 10.1109/TCYB.2020.2988093.
[41] C. Xu, H. Wang, X. H. Liu, and W. D. Zhao, "Bi-attention network for bi-directional salient object detection," *Appl. Intell.*, vol. 53, no. 19, pp. 21500–21516, 2023. doi: 10.1007/s10489-023-04648-8.
[42] Q. Wu, P. C. Zhu, Z. L. Chai, and G. D. Guo, "Joint learning of foreground, background and edge for salient object detection," *Comput. Vis. Image Underst.*, vol. 240, 2024, Art. no. 103915. doi: 10.1016/j.cviu.2023.103915.
[43] Y. -H. Wu, Y. Liu, L. Zhang, M. -M. Cheng, and B. Ren, "EDN: Salient object detection via extremely-downsampled network," *IEEE Trans. Image Process.*, vol. 31, pp. 3125–3136, 2022. doi: 10.1109/TIP.2022.3164550.

[44] Y. B. Han, L. J. Wang, S. L. Cheng, Y. M. Li, and A. Y. Du, "Residual dense collaborative network for salient object detection," *IET Image Process.*, vol. 17, no. 2, pp. 492–504, 2023. doi: 10.1049/ipr2.12649.
[45] C. Xu, X. H. Liu, and W. D. Zhao, "Attention-guided salient object detection using autoencoder regularization," *Appl. Intell.*, vol. 53, no. 6, pp. 6481–6495, 2023. doi: 10.1007/s10489-022-03917-2.
[46] X. F. Li, Y. Wang, T. Z. Wang, and R. L. Wang, "Spatial frequency enhanced salient object detection," *Inf. Sci.*, vol. 647, 2023, Art. no. 119460. doi: 10.1016/j.ins.2023.119460.
[47] L. Q. Zhang and Q. Zhang, "Salient object detection with edge-guided learning and specific aggregation," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 34, no. 1, pp. 534–548, 2024. doi: 10.1109/TCSVT.2023.3287167.
[48] J. Y. Zhu, X. B. Qin, and A. Elsaddik, "DC-Net: Divide-and-conquer for salient object detection," *Pattern Recognit.*, vol. 157, 2024, Art. no. 110903.
[49] Y. Lee, M. Lee, S. Cho, and S. Lee, "Adaptive graph convolution module for salient object detection," in *Proc. ICIP*, Kuala Lumpur, Malaysia, 2023, pp. 1395–1399.
[50] Y. Bi, Z. X. Chen, C. Y. Liu, T. Liang, and F. Zheng, "Supervised contrastive learning with multi-scale interaction and integrity learning for salient object detection," *Mach. Vision Appl.*, vol. 35, no. 4, 2024, Art. no. 74. doi: 10.1007/s00138-024-01552-0.
[51] G. Zhu, J. B. Li, and Y. H. Guo, "PriorNet: Two deep prior cues for salient object detection," *IEEE Trans. Multimedia*, vol. 26, pp. 5523–5535, 2024. doi: 10.1109/TMM.2023.3335884.
[52] X. Li, C. A. Wang, D. Ma, and X. Q. Wu, "Feature refinement from multiple perspectives for high performance salient object detection," in *Proc. PRCV*, Xiamen, China, 2024, vol. 14436, pp. 56–67.
[53] Y. Wang, R. L. Wang, X. Fan, T. Z. Wang, and X. J. He, "Pixels, regions, and objects: Multiple enhancement for salient object detection," in *Proc. CVPR*, Vancouver, BC, Canada, 2023, pp. 10031–10040.
[54] Y. G. Yi, N. Y. Zhang, W. Zhou, Y. J. Shi, G. S. Xie, and J. Z. Wang, "GPONet: A two-stream gated progressive optimization network for salient object detection," *Pattern Recognit.*, vol. 150, 2024, Art. no. 110330. doi: 10.1016/j.patcog.2024.110330.
[55] K. M. He, X. Y. Zhang, S. Q. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. CVPR*, Las Vegas, NV, USA, 2016, pp. 770–778.
[56] Z. L. Huang *et al.*, "CCNet: Criss-cross attention for semantic segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 6, pp. 6896–6908, 2023. doi: 10.1109/TPAMI.2020.3007032.
[57] W. Xu and Y. Wan, "ELA: Efficient local attention for deep convolutional neural networks," 2024, *arXiv:2403.01123*.
[58] L. -C. Chen, G. Papandreou, F. Schroff, and H. Adam, "Rethinking atrous convolution for semantic image segmentation," 2017, *arXiv:1706.05587*.
[59] Q. Yan, L. Xu, J. P. Shi, and J. Y. Jia, "Hierarchical saliency detection," in *Proc. CVPR*, Portland, OR, USA, 2013, pp. 1155–1162.
[60] L. J. Wang *et al.*, "Learning to detect salient objects with image-level supervision," in *Proc. CVPR*, Honolulu, HI, USA, 2017, pp. 3796–3805.
[61] C. Yang, L. H. Zhang, H. C. Lu, X. Ruan, and M. -H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. CVPR*, Portland, OR, USA, 2013, pp. 3166–3173.
[62] G. B. Li and Y. Z. Yu, "Visual saliency based on multiscale deep features," in *Proc. CVPR*, Boston, MA, USA, 2015, pp. 5455–5463.
[63] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. CVPR*, Providence, RI, USA, 2012, pp. 733–740.
[64] D. -P. Fan, C. Gong, Y. Cao, B. Ren, M. -M. Cheng and A. Borji, "Enhanced-alignment measure for binary foreground map evaluation," in *Proc. IJCAI*, Stockholm, Sweden, 2018, pp. 698–704.
[65] M. -M. Cheng and D. -P. Fan, "Structure-measure: A new way to evaluate foreground maps," *Int. J. Comput. Vis.*, vol. 129, no. 9, pp. 2622–2638, 2021. doi: 10.1007/s11263-021-01490-8.
[66] R. Margolin, L. Zelnik-Manor, and A. Tal, "How to evaluate foreground maps?" in *Proc. CVPR*, Columbus, OH, USA, 2014, pp. 248–255.

[67] R. Achanta, S. Hemami, F. Estrada, and S. Süsstrunk, "Frequency-tuned salient region detection," in *Proc. CVPR*, Miami, FL, USA, 2009, pp. 1597–1604.
[68] J. Li, J. M. Su, C. Q. Xia, M. C. Ma, and Y. H. Tian, "Salient object detection with purificatory mechanism and structural similarity loss," *IEEE Trans. Image Process.*, vol. 30, pp. 6855–6868, 2021. doi: 10.1109/TIP.2021.3099405.
[69] M. -M. Cheng, S. -H. Gao, A. Borji, Y. -Q. Tan, Z. Lin, and M. Wang, "A highly efficient model to study the semantics of salient object detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 44, no. 11, pp. 8006–8021, 2022. doi: 10.1109/TPAMI.2021.3107956.
[70] G. Zhu, J. B. Li, and Y. H. Guo, "Supplement and suppression: Both boundary and nonboundary are helpful for salient object detection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 34, no. 9, pp. 6615–6627, 2023. doi: 10.1109/TNNLS.2021.3127959.
[71] Z. Wu, L. Su, and Q. M. Huang, "Stacked cross refinement network for edge-aware salient object detection," in *Proc. ICCV*, Seoul, Republic of Korea, 2019, pp. 7263–7272.
[72] Z. Wu, L. Su, and Q. M. Huang, "Cascaded partial decoder for fast and accurate salient object detection," in *Proc. CVPR*, Long Beach, CA, USA, 2019, pp. 3902–3911.
[73] H. Y. Mei *et al.*, "Exploring dense context for salient object detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 32, no. 3, pp. 1378–1389, 2022. doi: 10.1109/TCSVT.2021.3069848.
[74] H. J. Zhou, X. H. Xie, J. -H. Lai, Z. X. Chen, and L. X. Yang, "Interactive two-stream decoder for accurate and fast saliency detection," in *Proc. CVPR*, Seattle, WA, USA, 2020, pp. 9138–9147.
[75] J. M. Su, J. Li, Y. Zhang, C. Q. Xia, and Y. H. Tian, "Selectivity or invariance: Boundary-aware salient object detection," in *Proc. ICCV*, Seoul, Republic of Korea, 2019, pp. 3798–3807.