**J. Meas. Sci. Instrum.**, 2026, 17(2): 195-207. http://jmsi.publish.founderss.cn/
**DOI**: 10.62756/jmsi.1674-8042.2026017 jmsi@nuc.edu.cn

# Salient object detection method based on object integrity enhancement guided by edge information

QIU Haoqing[1,2], GE Hongwei[1,2*], LI Ting[2]

*1. Engineering Research Center of Intelligent Technology for Healthcare, Ministry of Education, Jiangnan University, Wuxi 214122, China;*
*2. School of Artificial Intelligence and Computer Science, Jiangnan University, Wuxi 214122, China*

***Corresponding author:** GE Hongwei (ghw8601@163.com)

**Received:** February 7, 2025 **Revised:** April 1, 2025 **Accepted:** May 12, 2025

**Abstract:** In the saliency object detection task, a salient object detection method based on object integrity enhancement guided edge information is proposed to address the problems of blurred edges and object incompleteness in recognition results. Firstly, the diversity feature extraction module was proposed to capture the features of complex and variable salient objects through various convolutional operations, thereby enriching the feature representation of the model. Then, the object integrity enhancement module was designed to process the initial fused multi-level features in parallel, and the integrity information of salient objects was further enhanced by exploring spatial and channel branches. Finally, the edge feature enhancement module was employed to use the deep edge prediction features to guide the feature map to pay more attention to the foreground and background region and edge information, and to improve the model's edge perception capability. Experiments on four public datasets, such as ECSSD and DUTS-TE, showed that the proposed algorithm achieved higher detection accuracy than other advanced algorithms in several metrics, such as $S$-measure and $F$-measure on DUTS-TE dataset were 0.859 and 0.895, respectively. The proposed algorithm demonstrated superior capability in the perception and refinement of salient object boundaries, further enhancing its robustness in complex scenes.

**Key words:** salient object detection; diversity feature extraction; attention mechanism; edge information; feature enhancement; multi-level auxiliary supervision; convolutional neural network

## 0 Introduction

Salient object detection (SOD) is a task in computer vision that aims to automatically identify and extract the regions or objects in an image that attracts the most attention from the human eye. Because it focuses only on regions of interest, salient detection has been extensively utilized in diverse fields, such as semantic segmentation[1], visual tracking[2].

In the past, SOD was mainly based on traditional method[3-4], which relied primarily on handcrafted feature extraction. Although these methods did not require collecting large amounts of samples to learn the features of salient objects, their generalization ability was relatively weak due to the limited expressive power of the extracted features. As a result, they were typically only effective in simple situation. In recent years, convolutional neural network (CNN) have demonstrated significant advantages in feature extraction, leading to the emergence of numerous CNN-based saliency object detection methods[5-9]. These methods use the rich semantic features extracted to recognize salient objects and achieve promising results. However, these methods also suffer from the following issues.

1) Salient objects are often diverse in type and vary in size, such as the person and airplane shown in Fig.1. Simple combinations of convolutional layers and pooling layers lack sufficient flexibility in feature capture and fusion, making it difficult to effectively handle the variability of salient objects. This results in incomplete detection of the objects.




**Fig. 1 Problem of variable scale**

2) Objects often have complex edges, such as the chair, animals, and vehicles shown in Fig.2. Although low-level features retain a significant amount of spatial structure and

© The Author(s) 2026. Published by Publishing Center of North University of China. The articles published in this open access journal are distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

regular edge information, the downsampling process may lead to the loss of this information due to the reduction in pixel count. This can lead to low-quality edges in the extracted saliency maps, potentially resulting in blurry edges and incomplete structures.




**Fig. 2 Problem of edge complexity**

To effectively address the issues of object integrity and complex edges, we proposed the edge-information guided object-integrity enhancement network (EGOENet) for salient object detection. First, the diversity feature extraction module, constructed using diversified convolution kernels extracted rich and diverse spatial feature information, improving the model's representational capability while enhancing its adaptability to multi-scale objects. Then, the object integrity enhancement module (OIEM) was used to enhance and fuse feature information in different channels and spatial dimensions in parallel, enabling the network to fully represent the complete information of salient objects while suppressing irrelevant information. Finally, the edge feature enhancement module utilized rich deep-edge maps to further refine the foreground and background regions as well as the edges of salient objects. Additionally, the decoupled prediction mechanism helped mitigate the issue of edge blurring caused by the loss of information in low-level features, which were rich in edge information during downsampling. Additionally, we further adopted an auxiliary supervision strategy to guide the mode's training, which improves detection performance.

# 1 Related work

## 1.1 Scale-based learning approaches for SOD

To effectively tackle the challenge of scale variation, many researchers have analyzed it from various perspectives. Zhang et al.[5] designed a framework that aggregated multi-level convolutional features and utilized hierarchical features to provide robust scale information. In addition, Zhang et al.[6] achieved selective fusion of multi-scale contextual information through a multi-path recurrent feedback model, effectively transmitting global semantic information to shallow layers, thus enhancing the model's overall performance. Zhao et al.[7] proposed to use the $F$-measure loss, which can generate more accurate saliency maps while avoiding the limitations associated with traditional threshold dependence.

## 1.2 Attention mechanism approaches for SOD

Recently, the attention mechanism has been introduced in the SOD field to improve the model's capability to recognize and localize salient objects. The core idea of these approaches is to dynamically allocate and adjust the importance of different regions in the image, allowing the model to concentrate more effectively on the salient object areas while ignoring irrelevant backgrounds. Liu et al.[8] designed a pixel-level contextual attention network that generated attention maps for the contextual region of each pixel and selectively focused on the contextual information that helped with the saliency prediction of the current pixel. Meanwhile, Fang et al.[9] designed a cascaded dual-attention module with a pooling layer to compress features while retaining the texture and background information of the image, thereby obtaining more global contextual information.

## 1.3 Edge refinement approaches for SOD

Early feature fusion methods typically treated shallow features as edge features. However, shallow features contain some background noise. It is not suitable to directly use them as edge features. To overcome this, recent studies have introduced new methods based on edge feature learning. For instance, Qin et al.[10] adopted a hybrid loss, which combined cross-entropy loss, structural similarity loss, and intersection-over-union loss to guide the model's focus on pixel-level, region-level, and global saliency information in a multi-level manner, thereby achieving more precise detection results. Zhou et al.[11] designed a dual-stream decoder, allowing different parts of the network to interact and exchange information during the decoding stage, improving the understanding and capture of salient object edge information.

# 2 EGOENet

As shown in Fig.3, our model is structurally divided into two components: encoder and decoder. The encoding stage is based on the modified backbone network ResNet-50[12], which mainly removes the last two fully connected layers. The decoding stage is composed of diversity feature extraction module (DFEM), OIEM, and edge feature enhancement module (EFEM). To be more specific, the given input image is first passed through the encoder for

feature extraction, and the generated features are denoted as $f_{bk}^{(i)}(i=1,2,3,4)$. Next, we further extract and enrich the feature information from the backbone network using a diversity feature extraction module, which employs multiple convolution strategies while adjusting the feature channels to 64. Then, we feed the output features from DFEM and the intermediate features from decoupled branch of EFEM into the OIEM to achieve cross-level feature fusion and enhance the potential integrity information of the salient object. Finally, the edge feature enhancement module uses deep-layer ground truth prediction maps to guide the aggregated features from OIEM to strengthen the foreground and background regions, as well as edge information of the salient object.

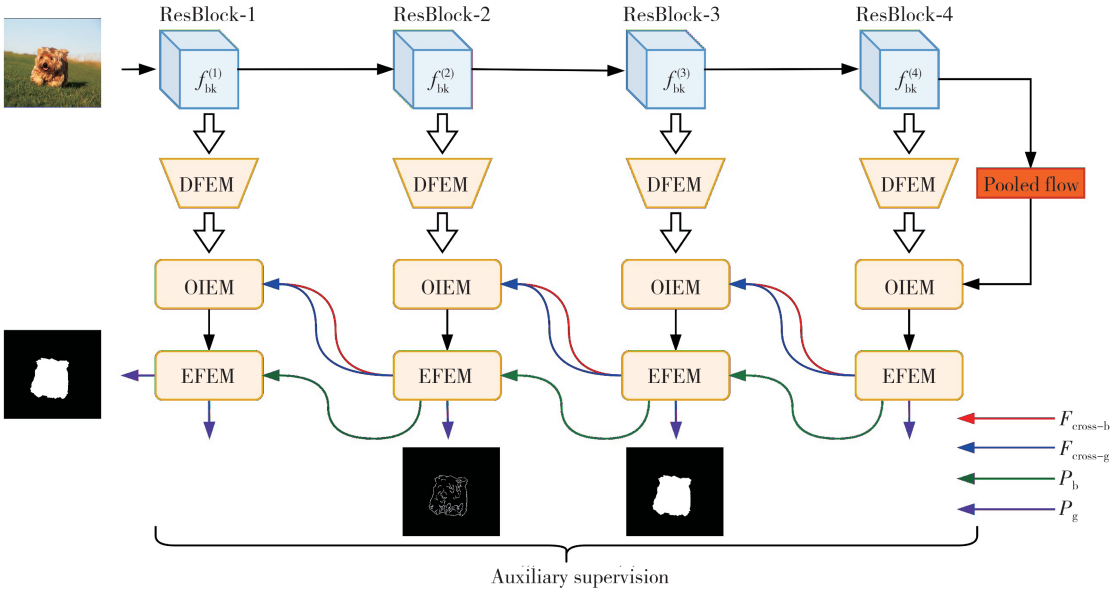



**Fig. 3 Overall architecture of network**

## 2.1 Diversity feature extraction module

Chen et al.[13] have demonstrated that diverse convolutional kernels and different receptive fields can help the network capture features of objects of various shapes and sizes. Inspired by this, this paper constructs the DFEM using atrous convolution[13], asymmetric convolution[14], original convolution, and pooling operations. Since each type of convolutional kernel or pooling operation extracts features differently, the module not only extracts diverse features between objects, but also captures more global contextual information, and the specific structure shown in Fig.4. The overall process is

$$F_{dfe}^{(i)}=\mathrm{CBL}\left(X_{asy}\left(f_{bk}^{(i)}\right)+X_{atr}\left(f_{bk}^{(i)}\right)+\sum_{j=1}^{4}X_{ori}^{(j)}\left(f_{bk}^{(i)}\right)\right),\quad(1)$$

where $f_{bk}^{(i)}$ represents the features extracted from the $i$-th block of the backbone network; CBL represents an operation that combines $3\times3$ convolution, batch normalization, and Leaky ReLU; $X_{asy}$, $X_{atr}$, and $X_{ori}^{(j)}$ represent asymmetric convolution, atrous convolution, and original convolution, respectively; $F_{dfe}^{(i)}$ represents the diversity features extracted from the $i$-th layer. After processing by these submodules, the number of channels is adjusted to 64.

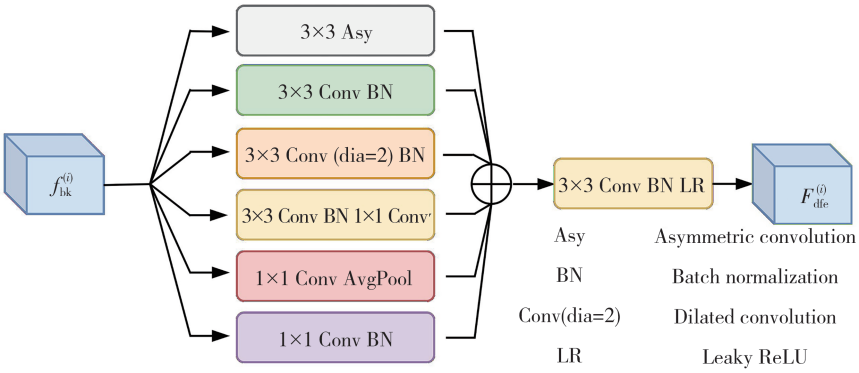



**Fig. 4 Illustration of DFEM**

For asymmetric convolution $X_{asy}$, it contains three levels of convolution kernels: a standard $3 \times 3$ kernel $K_{3\times3}$, a horizontal $1 \times 3$ kernel $K_{1\times3}$, and a vertical $3 \times 1$ kernel $K_{3\times1}$. All of them act on the same sliding window, and the details are

$$\boldsymbol{X}_{asy}(I) = \mathrm{SUM}(I*K_{3\times3}, I*K_{1\times3}, I*K_{3\times1}), \quad (2)$$

where $*$, SUM, and $I$ denote the 2D convolution, the element-wise addition, and the input feature, respectively.

The DFEM ultimately fuses the knowledge learnt from the asymmetric kernel, the atrous kernel, and the original kernel to enrich the feature space. Thus, it represents the important information in the image. We marked features generated by this module as $F_{dfe} = \left\{F_{dfe}^{(1)}, F_{dfe}^{(2)}, F_{dfe}^{(3)}, F_{dfe}^{(4)}\right\} \in \mathbf{R}^{64\times H\times W}$.

## 2.2 Object integrity enhancement module

Liu et al.[8] and Fang et al.[9] have achieved good results using channels or spatial attention mechanisms. They all essentially established correspondences between distinct features to emphasize the most important objects. However, how to fully explore the potential integrity feature of the object in both channel and spatial dimensions to reduce the phenomenon of missed detection and incorrect detection is still a problem worth studying. In this work, inspired by ICONet[15], we proposed the OIEM, and the structure is shown in Fig.5.

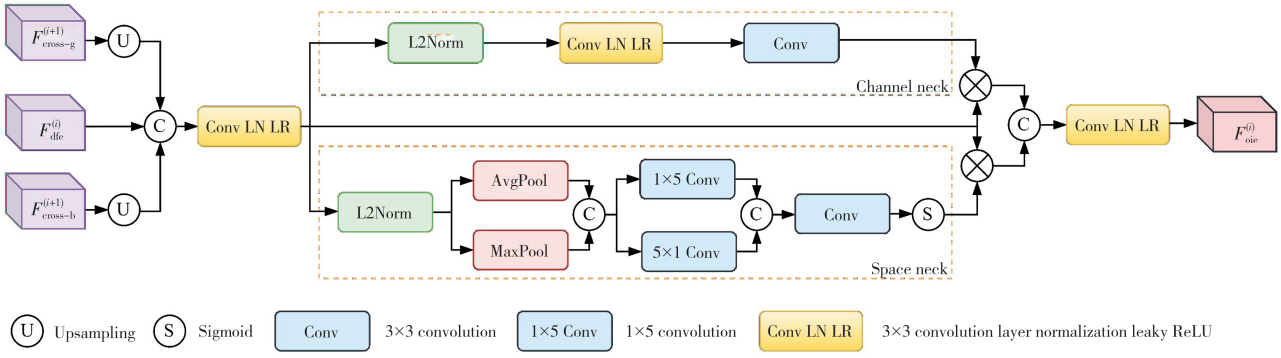



**Fig. 5 Illustration of OIEM**

This module applies a parallel mechanism to mine the relationships amongst channels and spatial dimensions, and further fuses this information, thereby enhancing the overall integrity of the salient object and effectively boosting the model's capability to understand the object structure.

Firstly, we adjust $F_{cross-b}^{(i+1)}$ and $F_{cross-g}^{(i+1)}$ to the same spatial resolution as $F_{dfe}^{(i)}$ using an upsampling operation. Then, the three features are concatenated to generate an initially enriched fused map $F_{fuse}^{(i)}$.

$$F_{fuse}^{(i)} = \mathrm{CLL}\left(\mathrm{Concat}\left[F_{cross-b}^{(i+1)}, F_{dfe}^{(i)}, F_{cross-g}^{(i+1)}\right]\right), \quad (3)$$

where CLL denotes the convolution operation with a kernel size of $3 \times 3$, followed by layer normalisation and Leaky ReLU.

Since the last layer of the OIEM doesn't have deep-layer inputs, we use pooled flow with rich semantic information to fill in the missing features, with specific details are shown in Fig.6.

After that, we parallelly mine the complete information of the feature $F_{fuse}^{(i)}$ in the channel and spatial dimension. In the channel dimension, $l_2$ norm is first performed to obtain the enhanced feature $I_{channel}^{(i)} \in [\mathrm{Bs}, C, 1, 1]$, which can effectively accelerate the convergence, reduce overfitting, and improve the model's stability through adjusting the weight decay. Then, to further enhance the integrity information between channels, we design the channel neck to learn feature $I_{channel}^{(i)}$. The design idea is

$$F_{channel}^{(i)} = F_{fuse}^{(i)} \otimes \mathrm{Conv}\left(\mathrm{CLL}\left(I_{channel}^{(i)}\right)\right), \quad (4)$$

where $\otimes$ denotes element-wise multiplication operation and Conv represents $3 \times 3$ convolution operation.

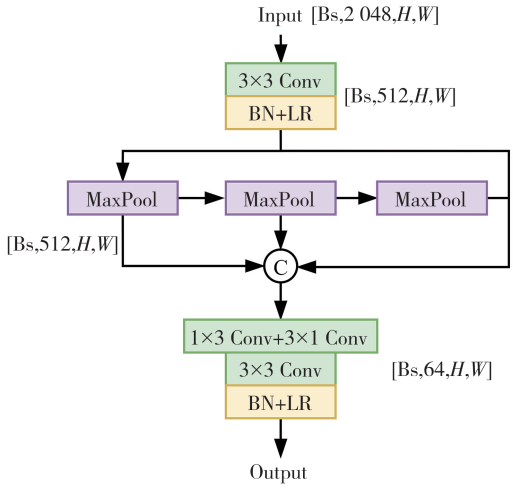



**Fig. 6 Illustration of pooled flow**

In the spatial dimension, $l_2$ norm is also performed to obtain the enhanced feature $I_{space}^{(i)} \in [\mathrm{Bs}, 1, H, W]$. Then, the enhanced feature is fed to the space neck to further enhance the integrity relationship between the spatial

dimensions. In this process, we use spatially separable convolution instead of original convolution, which allow us to retain the ability of learning feature in the original receptive field size under the condition of using fewer parameters. And by employing strip convolutions for directional modeling, it accurately captures the long-range dependencies and spatial structural information of the object. The specific design idea is

$$F=\mathrm{Concat}\left[\mathrm{AvgPool}\left(I_{\mathrm{space}}^{(i)}\right),\mathrm{MaxPool}\left(I_{\mathrm{space}}^{(i)}\right)\right], \quad (5)$$

$$F_{\mathrm{space}}^{(i)}=F_{\mathrm{fuse}}^{(i)}\otimes \sigma\left(\mathrm{Conv}\left(\mathrm{Concat}\left[\mathrm{Conv}_{1\times 5}(F),\mathrm{Conv}_{5\times 1}(F)\right]\right)\right), \quad (6)$$

where AvgPool and MaxPool represent the average and max pooling operations; $\sigma$ represents Sigmoid activation function; $\mathrm{Conv}_{1\times 5}$ and $\mathrm{Conv}_{5\times 1}$ mean convolutional operations with kernel size of $1\times 5$ and $5\times 1$, respectively.

Finally, we fuse the mined integrity information, with the design idea shown in Eq. (7). In addition, as shown in Fig. 7, after the features are input into the OIEM, the network is able to pay more attention to the integrity of the salient regions and distinguish between the foreground and background, which also demonstrates that the module has some ability to capture the integrity. We denote the features refined by OIEM as $F_{\mathrm{oie}}=\left\{F_{\mathrm{oie}}^{(1)},F_{\mathrm{oie}}^{(2)},F_{\mathrm{oie}}^{(3)},F_{\mathrm{oie}}^{(4)}\right\}\in\mathbf{R}^{64\times H\times W}$.

$$F_{\mathrm{oie}}^{(i)}=\mathrm{CLL}\left(\mathrm{Concat}\left[F_{\mathrm{channel}}^{(i)},F_{\mathrm{space}}^{(i)}\right]\right). \quad (7)$$

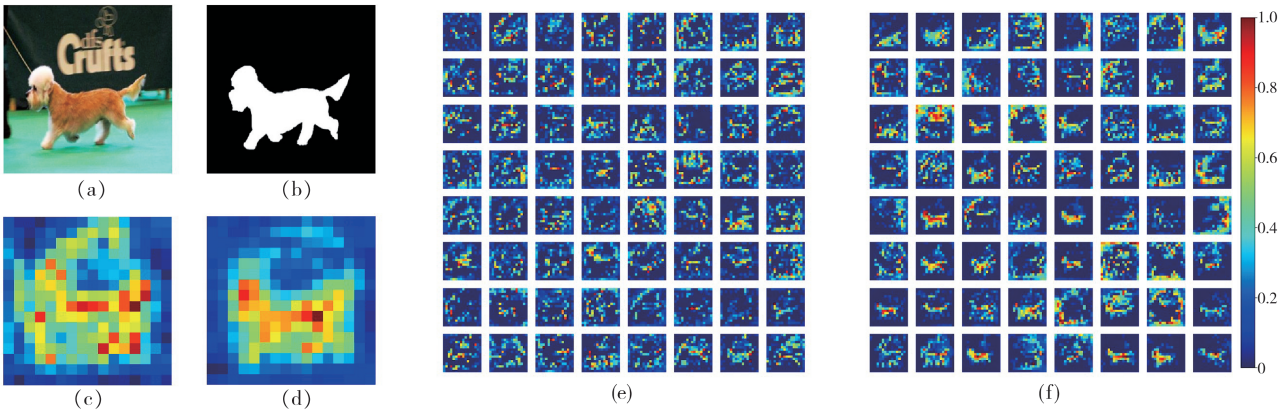



**Fig. 7 Visual comparison of heatmaps. (a) Input image; (b) Ground truth; (c) Feature heatmap before OIEM; (d) Feature heatmap after our OIEM; (e) Channel visualization before our OIEM; (f) Channel visualization after our OIEM**

## 2.3 Edge feature enhancement module

The key point of SOD lies in obtaining rich semantic features and accurate edge features. The former is used to distinguish salient objects from the background, while the latter facilitates the maintenance of clear edges of the salient object. The edge feature extraction in EGNet[16] is not directly related to the detection of salient regions, so the resulting edge features may differ from the salient objects. Inspired by this, we directly correlate the edge detection branch with the salient region detection branch, in order to simplify edge detection and capture accurate edge features. Therefore, we propose the EFEM, which is capable of detecting salient regions and edges simultaneously, with specific details shown in Fig. 8. *PR* is the PReLU activation function and *GN* is the group normalization. Additionally, EFEM processes the ground truth map by Canny algorithm[17] and generates object edge labels to supervise the training.

Specifically, we feed the feature layer $F_{\mathrm{oie}}^{(i)}$ and an edge map $P_{\mathrm{b}}^{(i+1)}$ into EFEM in parallel. The latter is utilized to guide the former to focus on edge information, as well as the foreground and background regions. Additionally, EFEM is implemented using a two-branch decoupling header.

For the foreground-background region branch, the first step is to subtract the edge map $P_{\mathrm{b}}^{(i+1)}$ from the feature map $F_{\mathrm{oie}}^{(i)}$ to obtain $F_{\mathrm{sub}}^{(i)}$. Then, we apply $F_{\mathrm{sub}}^{(i)}$ to the local attention (LA) and global attention (GA) branches, where the foreground and background features are further refined and enhanced using the attention mechanism. The LA branch uses $1\times 1$ convolution to scale the channels of the feature map, thereby enhancing the focus on local regions. While in the GA branch, global pooling is first used to retain global information across channels, and then a $1\times 1$ convolution is applied to capture long-range dependencies of the global information through compression and activation. Finally, the resulting feature coefficients are multiplied by

$$F_{\mathrm{sub}}^{(i)}=F_{\mathrm{cff}}^{(i)}-P_{\mathrm{b}}^{(i+1)}, \quad (8)$$

$$F_{\mathrm{sub}}^{(i)}=F_{\mathrm{sub}}^{(i)}\otimes\sigma\left(\mathrm{LA}\left(F_{\mathrm{sub}}^{(i)}\right)+\mathrm{GA}\left(F_{\mathrm{sub}}^{(i)}\right)\right). \quad (9)$$

Similarly, for edge information branch, the first step is to add the edge map $P_{\mathrm{b}}^{(i+1)}$ to the feature layer $F_{\mathrm{oie}}^{(i)}$ to

obtain $F_{\text{add}}^{(i)}$. Then, we feed $F_{\text{add}}^{(i)}$ into the LA and GA branches for learning. Finally, the resulting feature coefficients are multiplied by $F_{\text{add}}^{(i)}$. The formula is

$$F_{\text{add}}^{(i)} = F_{\text{cff}}^{(i)} + P_{\text{b}}^{(i+1)}, \tag{10}$$

$$F_{\text{add}}^{(i)} = F_{\text{add}}^{(i)} \otimes \sigma\left(\text{LA}\left(F_{\text{add}}^{(i)}\right) + \text{GA}\left(F_{\text{add}}^{(i)}\right)\right). \tag{11}$$

Eventually, the features from the two branches are fused using a $3 \times 3$ convolution to obtain $\text{Fuse}^{(i)}$. Next, the decoupled detection head process the feature $\text{Fuse}^{(i)}$ to generate an edge prediction map $P_{\text{b}}^{(i)}$ and an object prediction map $P_{\text{g}}^{(i)}$. For the last layer of EFEM, the feature map directly passes into the decoupled detection head for prediction. The edge prediction map and saliency prediction map after processing by EFEM are marked as $\left\{P_{\text{b}}^{(1)}, P_{\text{b}}^{(2)}, P_{\text{b}}^{(3)}, P_{\text{b}}^{(4)}\right\} \in \mathbf{R}^{1\times H\times W}$ and $\left\{P_{\text{g}}^{(1)}, P_{\text{g}}^{(2)}, P_{\text{g}}^{(3)}, P_{\text{g}}^{(4)}\right\} \in \mathbf{R}^{1\times H\times W}$, respectively.

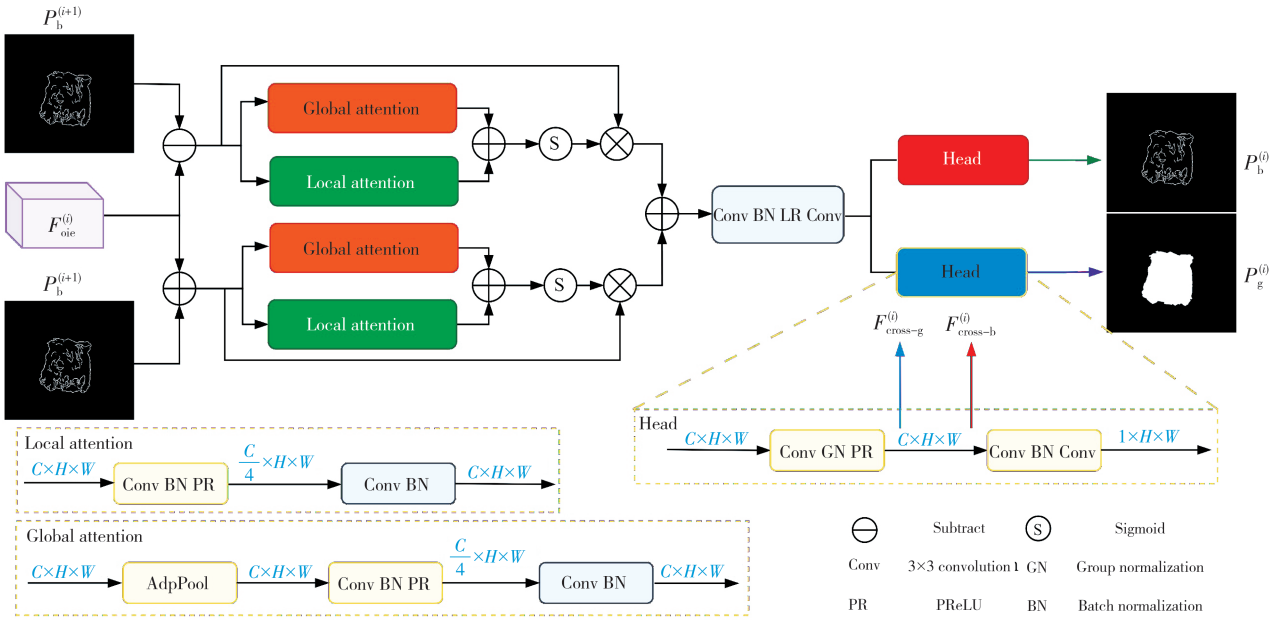



**Fig. 8 Illustration of EFEM**

## 2.3 Loss function

In the field of SOD, the binary cross-entropy (BCE) function has become the preferred loss function due to its effectiveness in pixel-level error metrics. It independently and equally calculates the error between the predicted map and the ground truth.

$$L_{\text{BCE}}(P,G) = -\sum_{(x,y)}\left[\begin{array}{l} G(r,c)\log_{\text{e}}\left(P(x,y)\right) + \\ \left(1-G(x,y)\right)\log_{\text{e}}\left(1-P(x,y)\right)\end{array}\right], \tag{15}$$

where $(x,y)$ denotes the pixel coordinates of the image; $P(x,y)$ and $G(x,y)$ represent the predicted map and the ground truth map at the location $(x,y)$. Additionally, the pixel values of $P(x,y)$ and $G(x,y)$ range from $[0,1]$.

The BCE loss function may be insufficient in accurately capturing the shape and edge of the object due to its equal weighting of foreground and background. To mitigate this issue, we introduce the intersection over union (IoU) loss function, which enhances the model's ability to recognise object structures by measuring the degree of overlap between the predicted region and the ground truth region. The formula is

$$L_{\text{IoU}}(P,G) = 1 - \frac{\sum_{(x,y)} G(x,y)P(x,y)}{\sum_{(x,y)}\left[G(x,y)+P(x,y)-G(x,y)P(x,y)\right]}. \tag{16}$$

Specifically, we employ a multi-level supervision strategy, which applies refined supervised learning to the saliency maps and edge saliency maps outputted by different layers of the network, facilitating the learning of more discriminative and robust saliency features. The overall loss function is defined as

$$L_{\text{g}}^{(i)} = L_{\text{BCE}}\left(P_{\text{g}}^{(i)}, G_{\text{g}}^{(i)}\right) + \mu L_{\text{IoU}}\left(P_{\text{g}}^{(i)}, G_{\text{g}}^{(i)}\right), \tag{17}$$

$$L_{\text{b}}^{(i)} = L_{\text{BCE}}\left(P_{\text{b}}^{(i)}, G_{\text{b}}^{(i)}\right) + \eta L_{\text{IoU}}\left(P_{\text{b}}^{(i)}, G_{\text{b}}^{(i)}\right), \tag{18}$$

$$L_{\text{total}} = \sum_{i=1}^{4}\left(L_{\text{g}}^{(i)} + L_{\text{b}}^{(i)}\right), \tag{19}$$

where $L_{\text{BCE}}$ and $L_{\text{IoU}}$ denote BCE loss and IoU loss, respectively; $\mu$ and $\eta$ take the values 3 and 0.1, respectively; $P_{\text{g}}^{(i)}$ and $P_{\text{b}}^{(i)}$ represent the predicted saliency map and edge saliency map from the $i$-th layer, respectively; $G_{\text{g}}^{(i)}$ and $G_{\text{b}}^{(i)}$ represent the ground-truth map and edge ground-truth

map, respectively.

## 3 Experiments

### 3.1 Datasets

Commonly used salient object detection datasets include DUTS[18], ECSSD[19], PASCAL-S[20], and HKU-IS[21]. The DUTS dataset covers different scenes and environments, consisting of 10 553 training images and 5 019 test images. The ECSSD dataset comprises 1 000 images, which are known for their rich semantic information and complex visual content. The PASCAL-S dataset consists of 850 highly challenging images with complex object relationships and backgrounds. HKU-IS consists of 4 447 images, the majority of which are characterised by overlapping boundaries or low contrast.

### 3.2 Experiment setup

In this work, we used the PyTorch 1.11.1 platform to train and test the model, with a 2080Ti GPU (11 GB memory) and a system equipped with 43 GB of RAM. In the data augmentation stage, random flipping, cropping, and scaling were introduced. We initialized the backbone network with the pre-trained ResNet-50 model on ImageNet during training period. The SGD optimizer with linear decay strategy was employed for auxiliary training. The learning rate of the backbone network was set to 0.001, while other parts were set to 0.01. The batch size and total number of training epochs were set to 32 and 100, respectively. The model achieved convergence after about 3.5 hours of training. In the prediction stage, the input images were uniformly resized to a resolution of $352 \times 352$ and then fed into the model to generate the saliency maps.

### 3.3 Evaluation metric

We utilized five widely adopted evaluation metrics to systematically evaluate the effectiveness of the proposed algorithm.

1) Mean absolute error (MAE) denotes the average of the absolute error per pixel between the predicted map and the ground truth. The calculation formula is

$$\mathrm{MAE}=\frac{1}{HW}\sum_{x=1}^{H}\sum_{y=1}^{W}\left|P(x,y)-G(x,y)\right|, \quad (20)$$

where $H$ and $W$ mean the height and width of the image, respectively.

2) $S$-measure ($S$) [22] mainly assesses the structural similarity between the predicted map and the ground truth. The design idea is

$$S=(1-\alpha)S_{\mathrm{r}}+\alpha S_{\mathrm{o}}, \quad (21)$$

where $S_{\mathrm{r}}$ denotes regional structural similarity; $S_{\mathrm{o}}$ denotes object structural similarity; $\alpha$ is set to 0.5.

3) PR curve visually demonstrates the model's performance balance at different confidence levels by calculating the precision and recall of the predicted map against the ground truth within the threshold range of [0, 255].

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \quad (22)$$

where $P$ represents precision; $R$ represents recall; TP, FP, and FN denote true positives, false positives, and false negatives, respectively.

4) $F$-measure ($F$) [23] is a method of comprehensively assessing model precision and recall in a weighted manner, which is computed by

$$F_{\beta}=\frac{(1+\beta^2)PR}{\beta^2(P+R)}, \quad (23)$$

where $\beta$ is set to 0.3[23].

5) $E$-measure ($E$) is intended to provide a comprehensive assessment of the overall similarity between saliency maps[24].

### 3.4 Comparison and analysis

To evaluate the algorithm proposed in this paper, we compared it with 10 recent advanced methods, including PiCANet[8], BASNet[10], CPD[25], $\mathrm{F^3Net}$[26], MINet[27], BFPINet[28], DCENet[29], RCSB[30], PoolNet+[31], and ICONet[15]. To ensure fairness of comparison, the same assessment code was used for all methods.

3.4.1 Quantitative evaluation

The network has capability to extract diverse features, deeply mine the comprehensive information of the object, and accurately capture edge details. The specific results are shown in Tables 1 and 2. The bolded values represent the optimal results; "↑" and "↓" favors larger and smaller values, respectively.

On the ECSSD dataset, EGOENet is 0.002 lower than the ICONet algorithm on $S$ score. But compared to other optimal methods, $F$, MAE, and $E$ scores are improved by 0.007, 0.002, and 0.003, respectively. On the DUTS-TE dataset, the model demonstrates optimal performance across all evaluation metrics, especially $E$ and $F$ scores, which are approximately 0.018 and 0.019 higher than RCSB, respectively. On the PASCAL-S dataset, except for the $S$ score, which is 0.001 lower than PoolNet+, $E$ score is increased by 0.008 and $F$ score is increased by 0.005. Benefiting from OIEM's enhancement of structural

consistency in salient object representation and EFEM's explicit edge supervision ensuring clear contours and spatial coherence, EGOENet achieves the highest performance metrics on the HKU-IS dataset, particularly in terms of $E$ and $F$ scores. These results not only validate the model's generalization ability in the salient object detection task but also demonstrate its adaptability and robustness to complex visual environments. Fig.9 shows the PR and $F$-measure curves with different methods obtained using four datasets, respectively.

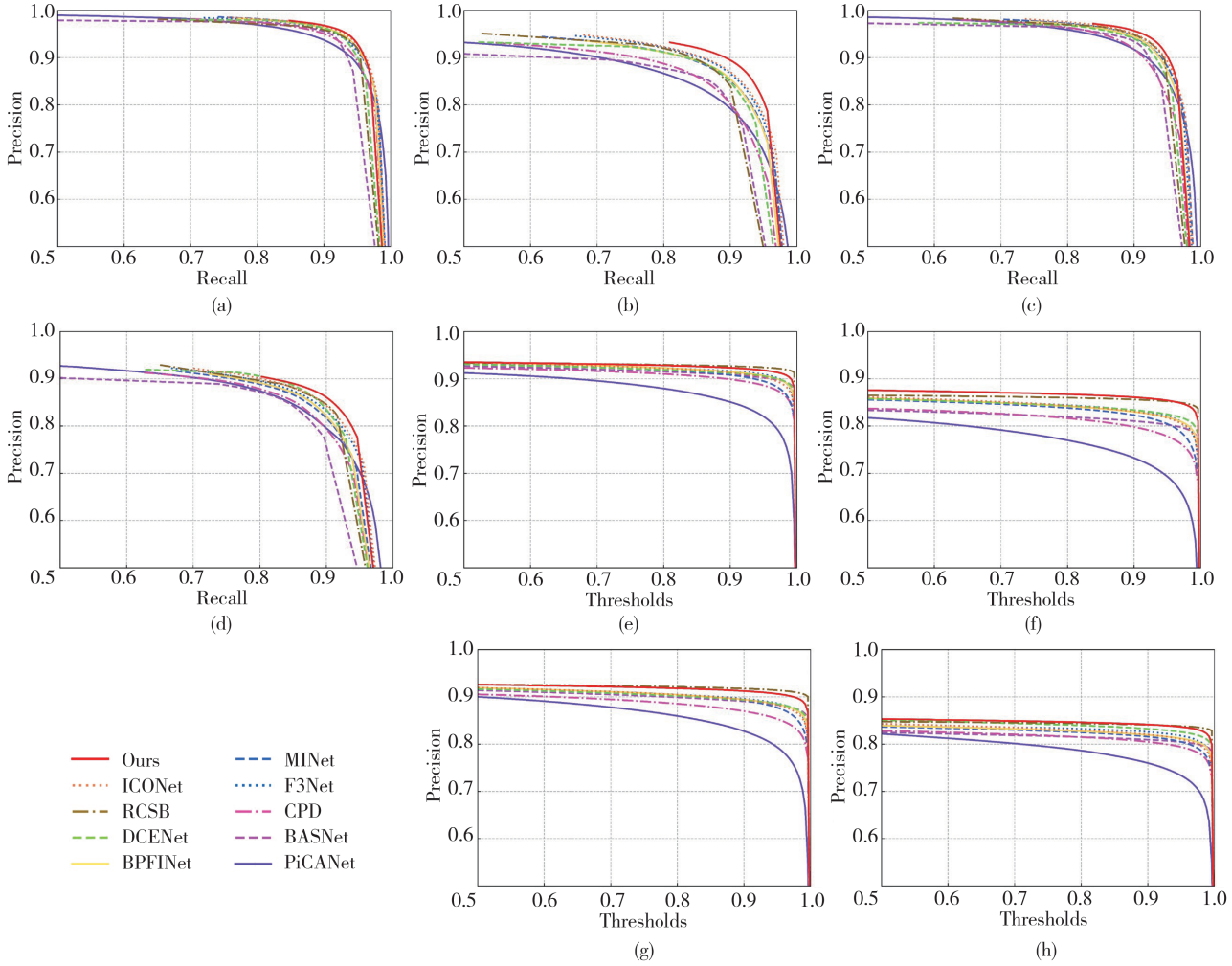



**Fig. 9 Curves of PR and $F$-measure. (a) PR curve of ECSSD; (b) PR curve of DUTS-TE; (c) PR curve of PASCAL-S; (d) PR curve of HKU-IS; (e) $F$-measure curve of ECSSD; (f) $F$-measure curve of DUTS-TE; (g) $F$-measure curve of PASCAL-S; (h) $F$-measure curve of HKU-IS**

A quantitative analysis of Fig.9 is conducted based on the evaluation metrics. As shown in Tables 1 and 2, the $E$ and $F$ scores of the proposed algorithm achieve optimal results across four datasets, demonstrating better overall performance and similarity. Furthermore, from Fig.10(a) to 10(h), it can be observed that the predicted maps of the proposed algorithm consistently outperform other detection algorithms in terms of overall structure.

3.4.2 Visual comparison

As shown in Fig.10(a) and 10(b), our method performed well in detecting salient objects at different scales. It indicates that under the DFE module, the algorithm can better adapt to variations in object scales while maintaining strong performance. As shown in Fig. 10(c) to 10(e), the text next to the dog and the chair are noise that is easily misclassified as background, while the bird's tail is a foreground object prone to misdetection. However, our method effectively distinguished and identified these objects. This demonstrates that the application of the OIEM significantly improves the algorithm's ability to capture the complete structure of the object, reducing both false negatives and false positives. Additionally, as observed from Fig.10(f) to 10(h), the application of the EFEM enables the algorithm to generate clearer and more accurate edges when handling complex edges, outperforming other saliency detection algorithms in terms of edge precision.

**Table 1 Quantitative comparison on ECSSD and DUTS-TE dataset**

| Method | ECSSD | | | | DUTS-TE | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE ↓ | $F$ ↑ | $S$ ↑ | $E$ ↑ | MAE ↓ | $F$ ↑ | $S$ ↑ | $E$ ↑ |
| PiCANet (TIP 17) | 0.046 | 0.865 | 0.914 | 0.923 | 0.054 | 0.747 | 0.861 | 0.872 |
| BASNet (CVPR 19) | 0.037 | 0.904 | 0.916 | 0.943 | 0.048 | 0.803 | 0.866 | 0.896 |
| CPD (CVPR 19) | 0.037 | 0.898 | 0.918 | 0.942 | 0.043 | 0.795 | 0.869 | 0.898 |
| F3Net (AAAI 20) | 0.033 | 0.912 | 0.924 | 0.948 | 0.035 | 0.835 | 0.888 | 0.920 |
| MINet (CVPR 20) | 0.033 | 0.911 | 0.925 | 0.950 | 0.037 | 0.825 | 0.884 | 0.917 |
| BPFI (Neurom. 21) | 0.034 | 0.915 | 0.926 | 0.951 | 0.038 | 0.831 | 0.886 | 0.917 |
| DCENet (TCSVT 22) | 0.035 | 0.913 | 0.921 | 0.949 | 0.038 | 0.834 | 0.882 | 0.918 |
| RCSB (WACV 22) | 0.034 | 0.916 | 0.922 | 0.950 | 0.035 | 0.840 | 0.881 | 0.919 |
| PoolNet+(TPAMI 23) | 0.040 | | 0.925 | | 0.039 | | 0.890 | |
| ICONet (TPAMI 23) | 0.032 | 0.918 | **0.929** | 0.954 | 0.037 | 0.837 | 0.889 | 0.924 |
| EGOENet (Ours) | **0.030** | **0.925** | 0.927 | **0.957** | **0.032** | **0.859** | **0.895** | **0.937** |

**Table 2 Quantitative comparison on PASCAL-S and HKU-IS dataset**

| Method | PASCAL-S | | | | HKU-IS | | | |
|---|---|---|---|---|---|---|---|---|
| | MAE ↓ | $F$ ↑ | $S$ ↑ | $E$ ↑ | MAE ↓ | $F$ ↑ | $S$ ↑ | $E$ ↑ |
| PiCANet (TIP 17) | 0.078 | 0.767 | 0.848 | 0.869 | 0.042 | 0.847 | 0.906 | 0.922 |
| BASNet (CVPR 19) | 0.076 | 0.793 | 0.838 | 0.879 | 0.032 | 0.889 | 0.909 | 0.943 |
| CPD (CVPR 19) | 0.071 | 0.794 | 0.848 | 0.882 | 0.034 | 0.875 | 0.905 | 0.938 |
| F3Net (AAAI 20) | 0.061 | 0.816 | 0.861 | 0.898 | 0.028 | 0.900 | 0.917 | 0.952 |
| MINet (CVPR 20) | 0.064 | 0.809 | 0.856 | 0.896 | 0.029 | 0.897 | 0.919 | 0.952 |
| BPFI (Neurom. 21) | 0.063 | 0.814 | 0.858 | 0.897 | 0.028 | 0.900 | 0.918 | 0.951 |
| DCENet (TCSVT 22) | 0.061 | 0.825 | 0.862 | 0.902 | 0.029 | 0.898 | 0.915 | 0.951 |
| RCSB (WACV 22) | **0.059** | 0.826 | 0.860 | 0.902 | 0.027 | 0.909 | 0.919 | 0.954 |
| PoolNet+(TPAMI 23) | 0.068 | | **0.864** | | 0.034 | | 0.921 | |
| ICONet (TPAMI 23) | 0.064 | 0.818 | 0.861 | 0.899 | 0.029 | 0.902 | 0.920 | 0.953 |
| EGOENet (Ours) | **0.059** | **0.831** | 0.863 | **0.910** | **0.025** | **0.916** | **0.923** | **0.961** |

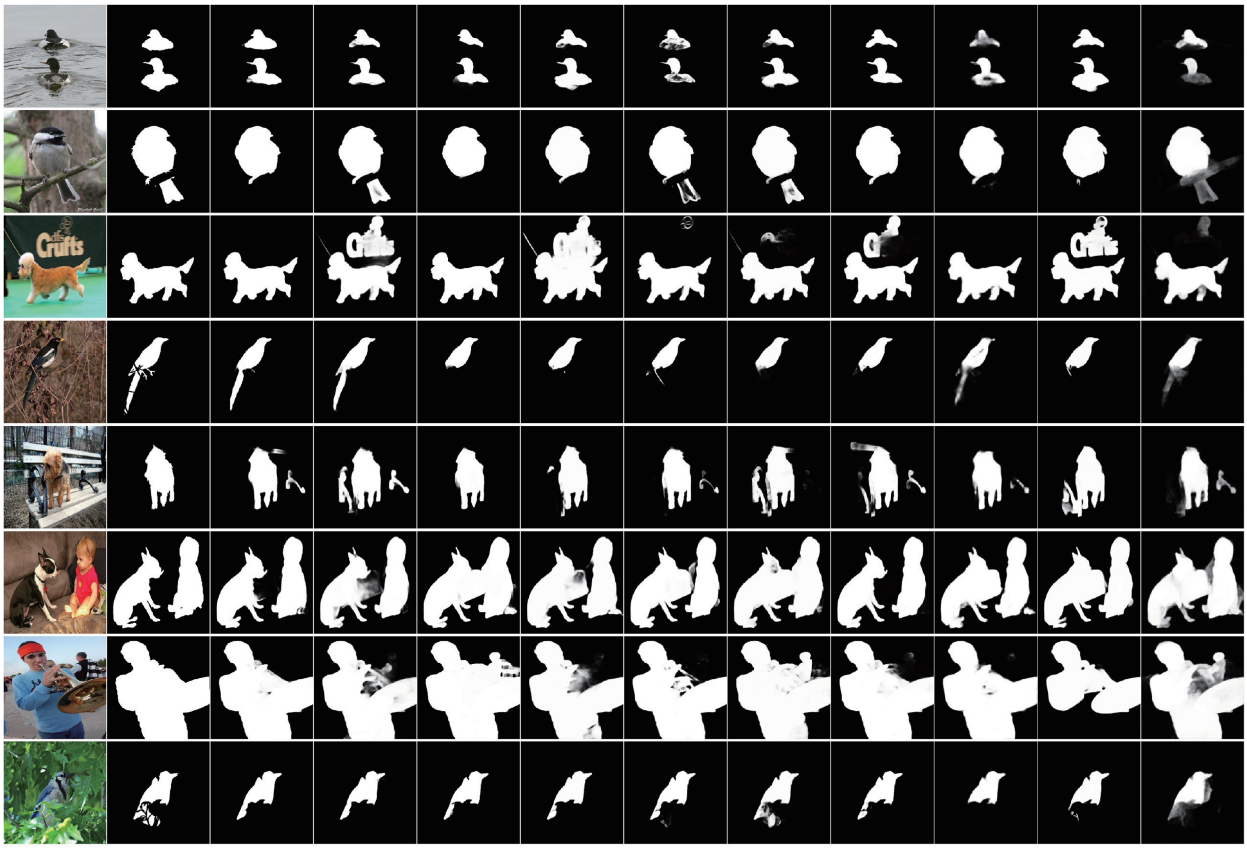



**Fig. 10 Visual comparison between proposed algorithm and other advanced algorithms**

## 3.5 Ablation experiment

### 3.5.1 Effectiveness of different components

We conducted ablation experiments on the DUTS-TE dataset to validate the effectiveness of the proposed module. In the experiments, the baseline network consists of a ResNet-50 encoder and a decoder composed of four VGG blocks. Then, the VGG blocks were replaced with the DFEM, followed by the sequential addition of the OIEM and the EFEM. The experimental results are shown in Table 3, and "√/×" indicate presence/absence of module. Compared to the baseline network, after incorporating the DFE module, the model shows a certain improvement in all evaluation metrics. This enhancement is attributed to the module's ability to extract more comprehensive features under different convolution kernels and its improved capability to adapt to multiple scales. Notably, $F$, $S$, $E$ scores are increased by 0.01, 0.005, and 0.004, respectively. After incorporating the OIEM, the model shows significant improvement in all evaluation metrics. This enhancement is attributed to the local details provided by the spatial branch and the global context offered by the channel branch, which together enables the model to more accurately capture the structural integrity of the object. The MAE score is decreased to 0.036, while the $F$ score is increased by 0.009, and the $S$ score is improved by approximately 0.004. Finally, after adding the EFEM, the model further refines and optimizes the edge features based on deep edge maps and the decoupled branches. The MAE score is decreased by 0.004, while the $S$ and $E$ scores are increased by approximately 0.009, and the $F$ score is improved by 0.013. These results fully demonstrate the effectiveness of the proposed method in enhancing model detection performance.

**Table 3 Ablation study on different modules**

| DFEM | OIEM | EFEM | MAE ↓ | $F$ ↑ | $S$ ↑ | $E$ ↑ |
|---|---|---|---|---|---|---|
| × | × | × | 0.038 | 0.827 | 0.878 | 0.923 |
| √ | × | × | 0.037 | 0.837 | 0.883 | 0.927 |
| √ | √ | × | 0.036 | 0.846 | 0.887 | 0.927 |
| √ | √ | √ | **0.032** | **0.859** | **0.895** | **0.937** |

To further demonstrate the effectiveness of each module, this subsection also conducted ablation experiments with visual comparisons of the different modules in the model. It can be observed from Fig. 11 that there is structural incompleteness in the generated prediction map when the OIEM is not added.




**Fig. 11 Comparison of saliency detection results of different modules**

It indicates that the OIEM plays an important role in the enhancement of object completeness information, making the whole salient map more complete. After adding the EFEM, the model significantly improves its ability to learn the edge features of salient objects, enhancing the precision of the edges and resulting in clearer prediction maps. Fig.12 demonstrates the visualization comparison of multi-level assisted supervision.

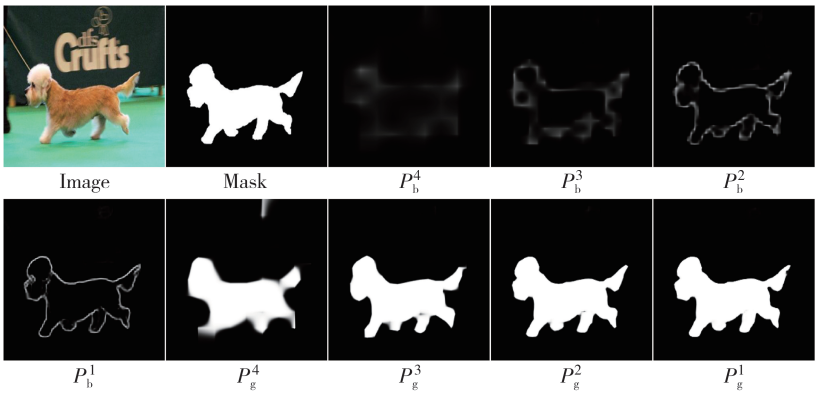



**Fig. 12 Visual comparison of multi-level complementary supervision**

It can be seen that the edge region of the edge prediction map and the object region of the saliency map are both a refinement process, proving the effectiveness of the multi-level assisted supervision model training.

3.5.2 Parameter tuning of loss function

To achieve better model performance, Table 4 presents the parameter tuning experiments for $\mu$ and $\eta$ in the loss function on the DUTS-TE dataset. Since the saliency map has a much greater overlap in terms of object structure compared to the edge map, the scaling factors of $\mu$ and $\eta$ are increased by factors of 10 during the experiment. By comparing the results with different values, the best model performance was achieved when $\mu = 3$ and $\eta = 0.1$.

**Table 4 Ablation study on parameter tunning**

| $\mu$ | $\eta$ | MAE ↓ | $F$ ↑ | $S$ ↑ | $E$ ↑ |
|---|---|---|---|---|---|
| 1 | 0.1 | 0.035 | 0.844 | 0.890 | 0.930 |
| 2 | 0.1 | 0.033 | 0.852 | 0.892 | 0.933 |
| **3** | **0.1** | **0.032** | **0.859** | **0.895** | **0.937** |
| 4 | 0.1 | 0.032 | 0.854 | 0.891 | 0.935 |
| 5 | 0.1 | 0.034 | 0.851 | 0.891 | 0.932 |

3.5.3 Ablation experiment of OIEM

The OIEM consists of spatial and channel branches. To further explore how to effectively organize the two branches to more comprehensively enhance the complete information of the salient object, we designed and evaluated the following validation experiment on the DUTS-TE dataset.

The experimental results are shown in Table 5.

**Table 5 Ablation study on OIEM combination method**

| Structure | MAE ↓ | $F$ ↑ | $S$ ↑ | $E$ ↑ |
|---|---|---|---|---|
| SBI+CH | 0.036 | 0.845 | 0.888 | 0.929 |
| SBI+DH | 0.033 | 0.850 | 0.891 | 0.932 |
| DBI+CH | 0.035 | 0.850 | 0.891 | 0.929 |
| DBI+DH | **0.032** | **0.859** | **0.895** | **0.937** |

The performance of the single-channel branch and single-spatial branch is slightly lower. Compared to the spatial branch-first cascade approach, the channel branch-first approach results in a 0.001 decrease in MAE, while the $F$ score changes from 0.850 to 0.853. Both the $S$ and $E$ scores are increased by approximately 0.006. When the spatial and channel branches are combined in parallel, the module benefits from the ability to simultaneously extract features from both spatial and channel dimensions, allowing different types of information to be processed and fused in parallel. This approach effectively avoids the potential information loss that may occur in the cascade method, thereby enhancing the model's comprehensive understanding of the salient object. Ultimately, the parallel structure enables the model to achieve the best performance, with a 0.001 decrease in MAE, and other metrics improving by approximately 0.005. This indicates that the parallel processing strategy is more effective in mining the structural integrity of the object, further enhancing the capability of salient object detection.

3.5.4 Ablation experiment of EFEM

To better evaluate the impact of the EFEM with the decoupling mechanism, the ablation experiments were conducted using the DUTS-TE dataset, specifically using the decoupling head (DH) and coupling head (CH). As shown in Table 6, the decoupled edge prediction branch enables the model to better capture edge details, while the object prediction branch provides accurate object location and shape. The EFEM decoupling head achieves better performance. The MAE score is decreased from 0.035 to 0.032, while the $F$ and $E$ scores are increased by 0.009 and 0.008, respectively, with the $S$ score improved by 0.004.

**Table 6 Ablation study on OIEM combination method**

| Structure | MAE ↓ | $F$ ↑ | $S$ ↑ | $E$ ↑ |
|---|---|---|---|---|
| Only channel | 0.034 | 0.849 | 0.889 | 0.931 |
| Only spatial | 0.034 | 0.851 | 0.885 | 0.932 |
| Channel+spatial cascade | 0.034 | 0.850 | 0.885 | 0.926 |
| Spatial+channel cascade | 0.033 | 0.853 | 0.891 | 0.933 |
| Spatial+channel parallel | **0.032** | **0.859** | **0.895** | **0.937** |

In addition, we also conducted ablation experiments on the dual-branch structure of the EFEM to validate its effectiveness, specifically using single-branch input (SBI) and dual-branch input (DBI). The introduction of the edge map effectively enhances the feature map's response in different regions. Whether using the decoupling head or the coupling head, the best performance is achieved when the dual-branch structure is adopted. Ultimately, the model achieves the highest results with the combination of DBI and DH.

## 4 Conclusions

A salient object detection method was proposed based on object integrity enhancement guided by edge information to tackle the problems of incomplete object structures and fuzzy edges in saliency detection. The method introduced a DFEM to capture feature diversity and adapt to multi-scale objects. The OIEM was employed to deeply mine the structural integrity of the salient object, reducing false negatives. Additionally, an EFEM was proposed to iteratively refine and correct the object edges using edge features, enhancing the model's edge perception. Extensive experiments demonstrate the effectiveness of our method. In future work, we will consider designing a more reasonable edge loss function to further enhance the network's ability to perceive and

refine object edges, and will also focus on applying the algorithm to practical engineering scenarios and generates meaningful impact.

## Acknowledgement

This work was supported by Priority Academic Program Development of Jiangsu Higher Education Institutions and the 111 Project (No. B12018).

## Declaration of conflicting interests

The authors have no conflict of interests related to this publication.

## References

[1] HOYER L, MUNOZ M, KATIYAR P, et al. Grid saliency for context explanations of semantic segmentation. Advances in Neural Information Processing Systems, 2019, 32: 6462-6473.

[2] ZHANG P, ZHUO T, HUANG W, et al. Online object tracking based on CNN with spatial-temporal saliency guided sampling. Neurocomputing, 2017, 257: 115-127.

[3] ZHOU Y, MAO A L, HUO S W, et al. Salient object detection via fuzzy theory and object-level enhancement. IEEE Transactions on Multimedia, 2019, 21(1): 74-85.

[4] LIN X, WANG Z J, MA L Z, et al. Saliency detection via multi-scale global cues. IEEE Transactions on Multimedia, 2019, 21(7): 1646-1659.

[5] ZHANG P P, WANG D, LU H C, et al. Amulet: aggregating multi-level convolutional features for salient object detection//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 202-211.

[6] ZHANG X N, WANG T T, QI J Q, et al. Progressive attention guided recurrent network for salient object detection//2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 18-23, 2018, Salt Lake City, UT, USA. New York: IEEE, 2018: 714-722.

[7] ZHAO K, GAO S H, WANG W G, et al. Optimizing the F-measure for threshold-free salient object detection//2019 IEEE/CVF International Conference on Computer Vision, October 27-November 2, 2019. Seoul, Korea. New York: IEEE, 2019: 8849-8857.

[8] LIU N, HAN J W, YANG M H. PiCANet: pixel-wise contextual attention learning for accurate saliency detection. IEEE Transactions on Image Processing, 2020, 29: 6438-6451.

[9] FANG J S, TAO Y H, ZHU G P. Saliency detection from complex background with attention-based boundary-aware pyramid pooling network. SSRN Electronic Journal, 2022: 4098532.

[10] QIN X B, ZHANG Z C, HUANG C Y, et al. BASNet: boundary-aware salient object detection//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2020: 7471-7481.

[11] ZHOU H J, XIE X H, LAI J H, et al. Interactive two-stream decoder for accurate and fast saliency detection//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020. Seattle, WA, USA. New York: IEEE, 2020: 9138-9147.

[12] HE K M, ZHANG X Y, REN S Q, et al. Deep residual learning for image recognition//2016 IEEE Conference on Computer Vision and Pattern Recognition, June 27-30, 2016, Las Vegas, NV, USA. New York: IEEE, 2016: 770-778.

[13] CHEN L C, PAPANDREOU G, SCHROFF F, et al. Rethinking atrous convolution for semantic image segmentation. 2017: arXiv: 1706.05587.

[14] DING X H, GUO Y C, DING G G, et al. ACNet: strengthening the kernel skeletons for powerful CNN *via* asymmetric convolution blocks//2019 IEEE/CVF International Conference on Computer Vision, October 27-November 2, 2019. Seoul, Korea. New York: IEEE, 2019: 1911-1920.

[15] ZHUGE M C, FAN D P, LIU N, et al. Salient object detection *via* integrity learning. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(3): 3738-3752.

[16] ZHAO J X, LIU J J, FAN D P, et al. EGNet: edge guidance network for salient object detection//2019 IEEE/CVF International Conference on Computer Vision, October 27-November 2, 2019. Seoul, Korea. New York: IEEE, 2019: 8779-8788.

[17] CANNY J. A computational approach to edge detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 1986, 8(6): 679-698.

[18] WANG L J, LU H C, WANG Y F, et al. Learning to detect salient objects with image-level supervision//2017 IEEE Conference on Computer Vision and Pattern Recognition, July 21-26, 2017, Honolulu, HI, USA. New York: IEEE, 2017: 3796-3805.

[19] YAN Q, XU L, SHI J P, et al. Hierarchical saliency detection//2013 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2013, Portland, OR, USA. New York: IEEE, 2013: 1155-1162.

[20] LI Y, HOU X D, KOCH C, et al. The secrets of salient object segmentation//2014 IEEE Conference on Computer Vision and Pattern Recognition, June 23-28, 2014, Columbus, OH, USA. New York: IEEE, 2014: 280-287.

[21] LI G B, YU Y Z. Visual saliency based on multiscale deep features//2015 IEEE Conference on Computer Vision and Pattern Recognition, June 7-12, 2015, Boston, MA, USA. New York: IEEE, 2015: 5455-5463.

[22] FAN D P, CHENG M M, LIU Y, et al. Structure-measure: a new way to evaluate foreground maps//2017 IEEE International Conference on Computer Vision, October 22-29, 2017, Venice, Italy. New York: IEEE, 2017: 4558-4567.

[23] ACHANTA R, HEMAMI S, ESTRADA F, et al. Frequency-tuned salient region detection//2009 IEEE Conference on Computer Vision and Pattern Recognition, June 20-25, 2009, Miami, FL, USA. New York: IEEE, 2009: 1597-1604.

[24] FAN D P, GONG C, CAO Y, et al. Enhanced-alignment measure for binary foreground map evaluation. 2018: arXiv: 1805.10421.

[25] WU Z, SU L, HUANG Q M. Cascaded partial decoder for fast and accurate salient object detection//2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 15-20, 2019, Long Beach, CA, USA. New York: IEEE, 2020: 3902-3911.

[26] WEI J, WANG S H, HUANG Q M. $F^3$Net: fusion, feedback and focus for salient object detection. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(7): 12321-12328.

[27] PANG Y W, ZHAO X Q, ZHANG L H, et al. Multi-scale interactive network for salient object detection//2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition, June 13-19, 2020. Seattle, WA, USA. New York: IEEE, 2020: 9410-9419.

[28] CHEN T Y, HU X G, XIAO J, et al. BPFINet: Boundary-aware progressive feature integration network for salient object detection. Neurocomputing, 2021, 451: 152-166.

[29] MEI H Y, LIU Y Y, WEI Z Q, et al. Exploring dense context for salient object detection. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(3): 1378-1389.

[30] KE Y Y, TSUBONO T. Recursive contour-saliency blending network for accurate salient object detection//2022 IEEE/CVF Winter Conference on Applications of Computer Vision, January 3-8, 2022, Waikoloa, HI, USA. New York: IEEE, 2022: 1360-1370.

[31] LIU J J, HOU Q B, LIU Z A, et al. PoolNet: exploring the potential of pooling for salient object detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2023, 45(1): 887-904.

# 边缘信息引导目标完整性增强的显著性目标检测算法

邱浩清[1,2], 葛洪伟[1,2*], 李　婷[2]

1. 江南大学 康养智能化技术教育部工程研究中心, 江苏 无锡 214122;

2. 江南大学 人工智能与计算机学院, 江苏 无锡 214122

**摘　要:** 在显著性目标检测任务中, 针对识别结果边缘模糊和目标不完整性的问题, 本文提出了一种边缘信息引导目标完整性增强的显著性目标检测算法。首先, 提出了多样性特征提取模块, 通过多种卷积核操作捕获复杂多变的显著物体的特征, 进而丰富模型的特征表达。然后, 设计了目标完整性增强模块, 并行地处理初步融合的多级特征, 并借助空间和通道探索分支进一步增强显著目标的完整性信息。最后, 提出了边缘特征增强模块, 利用深层的边缘预测特征引导特征图更多地关注前背景区域和边缘信息, 从而增强模型对边界的感知能力。在ECSSD、DUTS-TE等四个公开数据集上的实验表明, 所提算法在多个指标上较其它先进算法取得了较高的检测精度。其中, 在DUTS-TE上$S$-measure和$F$-measure的指标分别是0.859和0.895。所提算法在显著目标边缘的感知与优化方面表现出更优越的能力, 进一步提升了其在复杂场景的鲁棒性。

**关键词:** 显著性目标检测; 多样性特征提取; 注意力机制; 边缘信息; 特征增强; 多级辅助监督; 卷积神经网络

**引用格式:** QIU Haoqing, GE Hongwei, LI Ting. Salient object detection method based on object integrity enhancement guided by edge information. Journal of Measurement Science and Instrumentation, 2026, 17(2): 195-207. DOI: 10.62756/jmsi.1674-8042.2026017