IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 4 August 2023, accepted 17 September 2023, date of publication 22 September 2023,
date of current version 27 September 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3317901

RESEARCH ARTICLE

Multi-Scale Feature Enhancement for Saliency Object Detection Algorithm

SU LI[1], RUGANG WANG[1], FENG ZHOU[1], YUANYUAN WANG[1], AND NAIHONG GUO[2]
[1]School of Information Technology, Yancheng Institute of Technology, Yancheng 224051, China
[2]Yancheng XiongYing Precision Machinery Company Ltd., Yancheng 224006, China

Corresponding author: Rugang Wang (wrg3506@ycit.edu.cn)

This work was supported in part by the Jiangsu Graduate Practical Innovation Project under Grant SJCX22_1685 and Grant SJCX21_1517, in part by the Major Project of Natural Science Research of Jiangsu Province Colleges and Universities under Grant 19KJA110002, in part by the Natural Science Foundation of China under Grant 61673108, in part by the Natural Science Research Project of Jiangsu University under Grant 18KJD510010 and Grant 19KJB510061, and in part by the Natural Science Foundation of Jiangsu Province under Grant BK20181050.

ABSTRACT Aimed at existing saliency object detection models with problems of front and back view misclassification and edge blur, this study proposes an algorithm with multi-scale feature enhancement. In this algorithm, the feature maps of salient objects are extracted using VGG16. Multi-scale Feature Fusion Module is added to enhance the detailed information of the second feature layer and the semantic information of the fifth feature layer, which effectively improves the characterization ability of the second feature layer on the edges of salient objects and the fifth feature layer on salient objects. Simultaneously, Feature Enhancement Fusion Module is added to achieve the full fusion of local detail information and global semantic information through layer-by-layer fusion from deep to shallow, which is used to obtain a feature map with complete feature information. Finally, a complete prediction map with clear edges is obtained by training the network model. The performance of the proposed algorithm is compared with six algorithms, Amulet, R3Net, PoolNet, MINet, PurNet, and NSAL, on the HKU-IS, ECSSD, DUT-OMRON, and DUTS-TE datasets. MAE (Mean Absolute Error) values were decreased by 0.011, 0.009, 0, −0.001, 0.001, 0.003. F-measure were improved by 0.037, 0.019, 0.013, 0.017, 0.015, 0.09. E-measure were improved by: null, −0.008, 0.003, 0.005, −0.014, 0.047. S-measure were improved by: 0.073, 0.041, 0.016, 0.021, 0.016, 0.101. Compared with existing algorithms, the proposed algorithm can obtain better detection results and accurately identify all regions of significant objects.

INDEX TERMS Salient object detection, multi-scale feature fusion, feature-enhanced, local and global information.

I. INTRODUCTION

Visual attention mechanism is a psychological regulation mechanism that plays a very important role in the visual information processing process. Using the principle of the visual attention mechanism, salient object detection (SOD) can quickly and accurately detect salient targets in images and videos and simultaneously highlight the most interesting areas in vision [1], [2], [3]. Currently, SOD is widely used in scenarios such as robot navigation [4], semantic segmentation [5], and object recognition and detection [6], [7], [8], [9], [10], [11], [12]. Currently, SOD algorithms can be divided into saliency prediction based on eye movement points, and SOD with accurate object contour information. Among them, SOD with accurate object contour information is divided into two types of methods: traditional methods and deep neural networks to extract semantic features.

The traditional method mainly involves segmenting the area of the image, extracting shallow features such as color, shape, and edges, and performing saliency calculation directly, or using algorithms to fuse the feature layers to obtain saliency feature maps [13], [14], [15]. In 2020, Cui et al. [13] proposed saliency object detection based on multiple features and prior information, which detects saliency objects with more complete edges through the fusion of multiple cues, such as contrast, color, and texture features.

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

However, this method still has room for improvement in the detection performance of non-salient regions in processing complex images. In 2021, Zhang et al. [14] proposed a salient object detection method based on texture and color features using a Gabor filter and Bayesian algorithm to obtain clear salient targets. However, this method is prone to problems of missed detection and misclassification, and the detection accuracy still has room for improvement. In 2022, Ouyang et al. [15] proposed a defect detection algorithm for complex texture ceramic tiles based on the visual attention mechanism, which mainly uses the contrast principle and high-frequency suppression principle to detect the background texture and color of tiles to detect surface defects in complex textured tiles. However, this method is prone to false detection problems for niche color types of tiles and the detection performance can be improved. Traditional saliency detection methods can achieve satisfactory results when dealing with simple scene images. However, due to the reason that such methods cannot extract the deep semantic features of images, they lead to a low detection accuracy when dealing with complex images.

Convolutional Neural Networks (CNNs) are widely used by researchers in salient target detection tasks because they can obtain deep feature information of salient targets, thus greatly improving the detection accuracy of the targets to be tested [16], [17], [18], [19]. In 2021, Pang et al. [16] proposed a Multi-scale Interactive Network (MI-Net) for SOD, which mainly used an aggregated interaction module, self-interaction module, and consistency enhancement loss to integrate features at adjacent levels, obtain efficient multi-scale features, and maintain intraclass consistency. The experimental results show that the network achieves good prediction results on five commonly used significant target detection datasets and achieves significant performance improvement. In 2022, Fang et al. [17] proposed saliency detection from a complex background using an Attention-based Boundary-aware Pyramid Pooling Network, which helps the network to better retain background and texture information by constructing cascaded dual attention modules, feature aggregation modules, and boundary-aware modules to obtain more accurate saliency maps. However, this method suffers from a more complex network structure and longer training time. In 2023, Wang et al. [18] proposed a salient object detection method based on multi-scale feature fusion guided by edge information (EGMFNet), which is mainly used to enhance spatial and edge features by building multi-channel fusion residual blocks and a global spatial attention module with edge information guidance to obtain a clearer saliency map of the edges. However, this method still exhibits improvement when dealing with bright light. In 2023, Yang et al. [19] proposed a Dual-Stream Fusion and Edge-Aware Network for Salient Object Detection, which mainly constructs a multi-scale channel interaction module, dual-stream aggregation module, and boundary perception structure to obtain a fine saliency map at the edges. However, this method is computationally intensive and can be enhanced in terms of multi-scale perception and semantic understanding.

Based on existing research results, significant object detection technology based on convolutional neural networks has made great progress in terms of detection accuracy and training time, but there are still elements that require further optimization. (1) When convolutional technology is used to process feature maps, feature information is lost during the transmission process, leading to problems such as blurred edges of a significant object. (2) In scenes with complex backgrounds, convolution alone cannot accurately separate salient targets from background features, leading to problems such as the low accuracy of salient target detection. (3) Multiple uses of convolutional techniques to extract different levels of feature information leads to poor correlation of the feature information, resulting in incomplete saliency objects.

To solve these problems, this study proposes multi-scale feature enhancement for a saliency object detection algorithm that extracts feature maps of saliency objects using VGG16. Simultaneously, multi-scale feature enhancement and feature enhancement fusion modules were added to the algorithm. By fusing information from deep to shallow layers layer-by-layer, full fusion of local detail information and global semantic information is achieved to obtain a feature map with complete feature information to improve the detection performance of the model.

II. SYSTEM ARCHITECTURE

The structure of the proposed multi-scale feature-enhanced SOD network is illustrated in Figure 1. From the figure, it can be observed that the model includes a Feature Extraction Module (FEM), Multi-scale Feature Fusion Module (MFFM), Feature Enhancement Fusion Module (FEFM) and prediction output module. First, the feature information of the salient object was selected using the VGG16 network, and five feature layers of $F(i)\{i = 1, 2, 3, 4, 5\}$ were selected as backbone features. Second, the MFEM is used to enhance the capability of layer2 to characterize the edges of salient objects and layer5 to characterize salient targets. To improve the performance of the MFEM, spatial attention was added to layer2. By enhancing the location and detailed information of salient objects, the model's ability to recognize the edges of salient target regions can be improved. Channel attention was added to layer5. By enhancing the semantic features of salient regions, the recognition ability of the model for salient object regions can be improved. The FEFM is then used to achieve layer-by-layer fusion of deep-to-shallow information. Furthermore, full fusion of local detail information and global semantic information was achieved to obtain a feature map with complete feature information. Finally, the prediction output module was used to train the network model and output the prediction map of the salient object with completely clear edges. The algorithm primarily improves the detection performance of the model by improving the salient object features in the feature layer and expanding the receptive field.

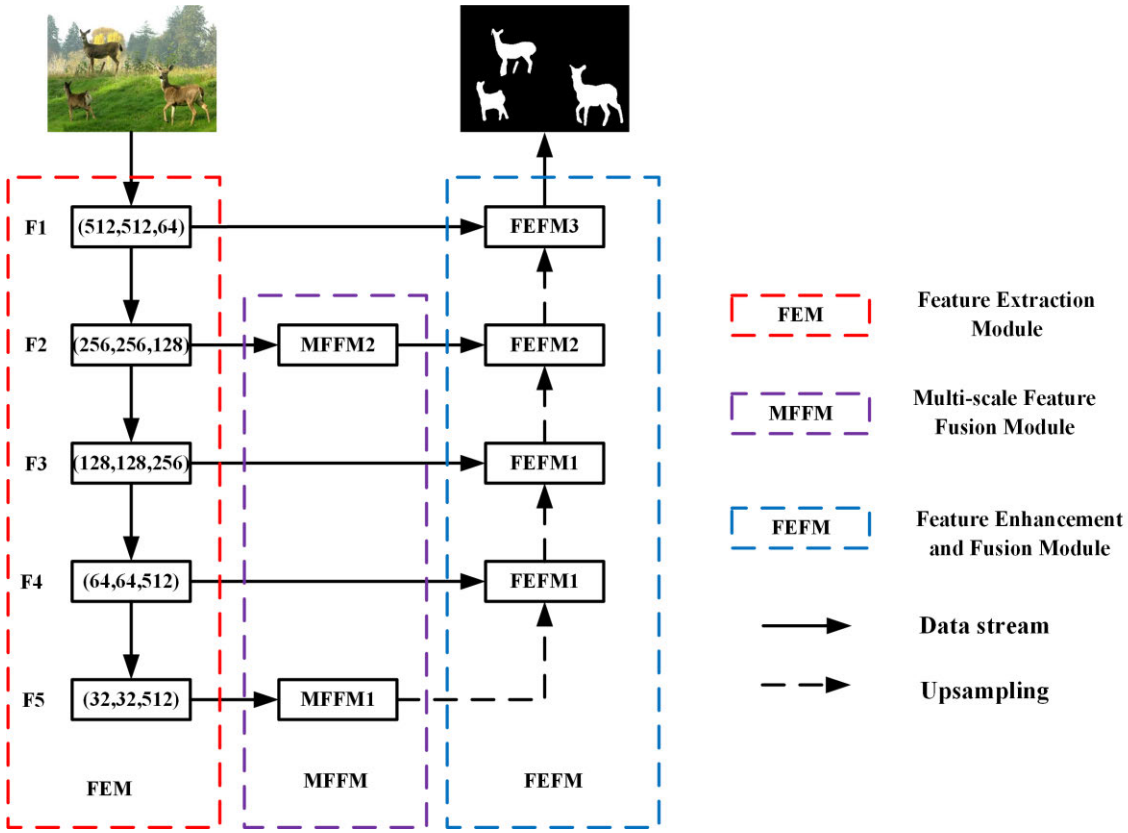


FIGURE 1. Network structure diagram.

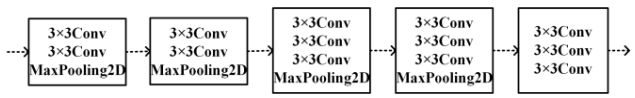


FIGURE 2. Feature extraction module.

A. FEATURE EXTRACTION MODULE

Compared with the complex structure of CNN and Transformers, VGG16 has a relatively simple structure, using only convolutional, pooling, and fully connected layers, which makes it easier to understand and implement. At the same time, VGG16 has strong feature extraction capability and is suitable for a wide range of image-related tasks. Therefore, VGG16 is used as the backbone network in this study. Meanwhile, in order to reduce the information loss of saliency targets, the extraction part is changed to a fully convolutional network. That is, the first 13 convolutional layers of the VGG16 network are retained, while the last two fully connected layers and the last pooling layer are removed, and the feature extraction module is shown in Figure 2. The specific implementation steps of the feature extraction stage are as follows: two 3 × 3 convolutions and one MaxPooling; two 3 × 3 convolutions and one MaxPooling; cubic 3 × 3 convolution and one MaxPooling; cubic 3 × 3 convolution and one MaxPooling; and cubic 3 × 3 convolution. Finally, five feature layers (F1, F2, F3, F4, and F5) were outputted, and the dimensions of the feature layers were 512 × 512 × 64. 256 × 256 × 128, 128 × 128 × 256, 64 × 64 × 512, and 32 × 32 × 512.

B. MULTI-SCALE FEATURE FUSION MODULE

In this study, Multi-scale Feature Fusion Module (MFFM) is added to enhance the ability of layer2 to characterize the significant object edges on the one hand, and layer5 to characterize the significant objects on the other hand. The structure of MFFM is shown in Figure 3. To enhance the feature information, MFFM was introduced in layer2 and layer5. Since layer2 contains rich positions and detailed information, it is more sensitive to spatial information. Therefore, the Spatial Attention Module (SAM) is used to enhance the location and detailed features of salient regions, thus improving the model's ability to recognize the edges of the salient target regions. Since layer5 contains rich semantic information, it is more sensitive to channel information. Therefore, the Improved Efficient Channel Attention (IECA) is used to enhance the semantic features of salient regions, thereby improving the model's ability to recognize salient object regions.

The overall flow of MFFM1 used in layer5 is shown in Figure 3(A). First, the feature layer was sampled in parallel using convolutions with different sampling rates, including 1 × 1, 3 × 3, and 5 × 5 convolutions. To reduce the computational and parametric sizes of the model, a 3 × 1 convolution plus 1 × 3 convolution was used instead

of a 3 × 3 convolution and a 5 × 1 convolution plus 1 × 5 convolution was used instead of a 5 × 5 convolution. Second, the obtained results are concatenated to increase the number of channels. The number of channels was adjusted using 1 × 1 convolution. Next, an improved version of efficient channel attention was used to increase the network's attention to the channel information of salient regions and improve the model's ability to identify salient object regions. The two results, which underwent parallel sampling and attention enhancement, were concatenated to increase the number of channels. Finally, information fusion and channel adaptation were performed based on the above results, using two 3 × 3 convolutions.

The overall flow of the MFFM2 used in layer 2 is shown in Figure 3(B). The parallel sampling procedure was the same as that described in Figure 3(A). In the second step, the results obtained in the first step were concatenated, and the number of channels was increased. The number of channels was adjusted using 1 × 1 convolution. Next, spatial attention is used to increase the network's attention to the spatial information of salient regions and improve the model's ability to identify the edges of the salient target regions. The two results, which underwent parallel sampling and attention enhancement, were concatenated to increase the number of channels. Finally, information fusion and channel adaptation are performed based on the above results using a 1 × 1 convolution.

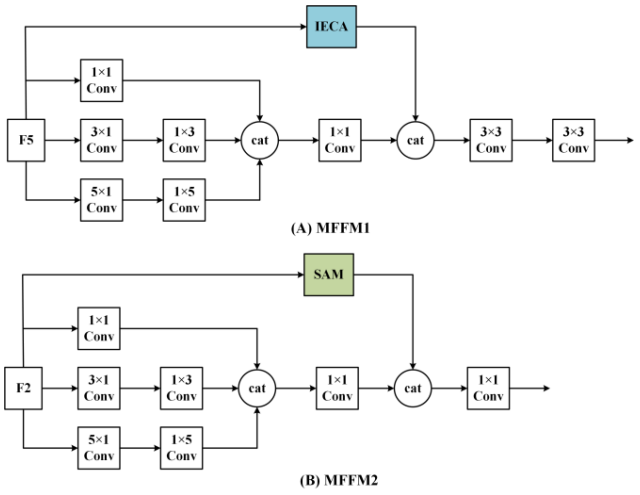


FIGURE 3. Multi-scale feature fusion module.

C. FEATURE ENHANCEMENT AND FUSION MODULE

In this study, a Feature Enhancement Fusion Module (FEFM) was added to enhance the information fusion of deep semantic features and shallow detailed features using up-sampling, residual structure, and concatenating operations to promote the fusion of global semantic information and local detailed information to highlight salient objects. The structure of FEFM is shown in Figure 4.

Because the shallow feature-layer receptive field has less overlap and a higher resolution, it can provide more detailed information, but has low semantic and high noise. The receptive field of the deeper feature layer overlaps more and has higher semantics, which can provide more global information but has a lower resolution and poor perception of detail. Therefore, enhanced fusion of shallow features and deep semantic information is an important component of saliency object detection.

The FEFM1 used for Layer4 is shown in Figure 4(A). The implementation steps were as follows: First, the feature layer obtained from F5 is u-sampled by MFFM to obtain P5_up, which has the same resolution as F4. Second, F4 and P5_up are concatenated to expand the channels. Next, a 1 × 1 convolution was used to adjust the number of channels to reduce the number of parameters and computation of the model. The Mixed Attention Module (MAM) is then used to further improve the location and detailed features of the salient objects. We then concatenated the results from the fourth step with F4 for the residuals to enrich feature-level information. Finally, two 3 × 3 convolutions were used to deepen the fusion of the feature information while adjusting the number of channels.

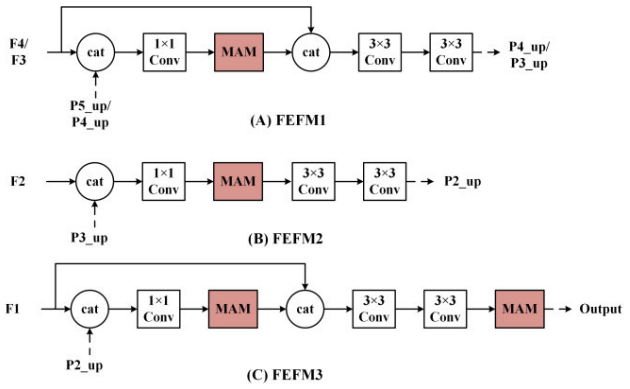


FIGURE 4. Feature enhancement fusion module.

The FEFM1 used for Layer3 is shown in Figure 4(A). The implementation steps are as follows. First, the result of the F4 processing is u-sampled to obtain P4_up, which has the same resolution as F3. Next, F3 and P4_up are concatenated to expand the channel. The next steps are the same as those in the back half of FEFM1.

During our experiments, we found that for the Layer2, whether to add jump connections in FEFM or not has no significant effect on the performance improvement of the algorithm and adds more computational overhead, which is unnecessary and not worth the cost for the algorithm design. Therefore, we eliminate the addition of jump connections in FEFM2. The FEFM2 used for the Layer2 is shown in Figure 4(B). The implementation steps are as follows. First, the result of F3 processing is u-sampled to obtain P3_up, which has the same resolution as F2. Second, the result of F2 after FEFM2 is concatenated with P3_up to expand the

number of channels. Subsequently, the number of channels was adjusted using a 1 × 1 convolution to reduce the number of parameters and computation of the model. The semantic features of salient objects were further enhanced using the MAM. Finally, the feature information is deeply fused using two 3 × 3 convolutions while adjusting the number of channels.

The FEFM3 used for Layer1 is shown in Figure 4(C). The implementation steps are as follows. First, the processed result of F2 is u-sampled to obtain P2_up, which has the same resolution as that of F1. Second, F1 and P2_up are concatenated to expand the channel. The steps are the same as those in the back half of FEFM1. Finally, the MAM was used again to further enhance the salient object features.

D. ATTENTION MODULE

In this study, the attention module was used to help the network model adaptively adjust the weights of the spatial and channel features in the feature layers of different scales, to reduce the position shift caused by the overlapping receptive field. This, in turn, reduces the negative impact of multiple convolutions and u-sampling, thus improving the network's ability to learn salient object features and facilitating the fusion of different feature information. Because the focus of feature-rich information in the shallow and deep feature layers is different, this study uses the Spatial Attention Module (SAM), Improved Efficient Channel Attention (IECA), and a combination of both, namely the Mixed Attention Module (MAM). The structure of the attention mechanism is shown in Figure 5.

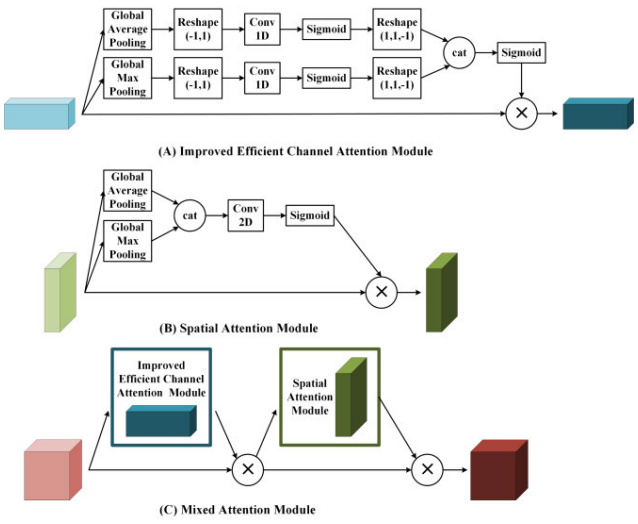


FIGURE 5. Structure of attention module.

The deep feature layer has more semantic feature information and is more sensitive to channel information; therefore, IECA was used to enhance the salient region features. The structure of IECA is shown in Figure 5(A). The Efficient Channel Attention (ECA) mentioned in the literature [20] only has a global average pooling operation, which can better preserve the background information but causes the image to be blurred. To solve this problem, referring to the Channel Attention Module (CAM) structure in the literature [21], a global maximum pooling operation was added in this study to enable the network to learn background information better and improve image clarity.

The shallow feature layer has more detailed feature information and is more sensitive to spatial information; therefore, SAM was used to enhance the salient region features. The SAM structure is shown in Figure 5(B).

Different feature layers have different emphases on feature information; therefore, conventional feature fusion does not reflect the correlation of spatial and channel features in different scale features, which leads to overlapping feature information and thus reduces model detection performance. To solve this problem, this study uses MAM to adaptively adjust the weight of spatial and channel features in different scale feature layers, which focuses on the correlation between channels as well as the spatial correlation of feature information, thus helping the network model to identify salient regions more accurately.

MAM consists of a tandem combination of IECA and SAM, and its structure is shown in Figure 5(C). The input feature layer enters the IECA first, and the output of the IECA is used as the input of the SAM, which not only allows the network model to focus on the salient object features first but also retains more contextual information, thus improving the detection performance of the network model.

III. RESULTS AND ANALYSIS

A. EXPERIMENTAL ENVIRONMENT AND DATASET

The experimental platform is NVIDIA GeForce RTX3090 GPU with 24GB memory. The network architecture in this study was based on the TensorFlow2.4. And we have used Python3.7 to complete the algorithm. The weights of the backbone network were obtained by pretraining on ImageNet. The stochastic gradient descent (SGD) algorithm was used in the experiments, the initial learning rate was set to 0.005, cosine annealing was used to adjust the learning rate of the model in the training process, and the momentum was set to 0.9. To prevent overfitting and promote convergence, the decay weight is set to 0.0005. Considering the computational volume and model convergence, the freeze training epoch was set to 50, thaw training epoch was set to 350, and batch size was set to 8.

The datasets used in this study were the ECSSD, DUT-OMRON, DUTS-TR, DUTS-TE, and HKU-IS. 1000 images in the ECSSD dataset, 5168 images in the DUT-OMROM dataset, 10553 images in the DUTS-TR dataset, 5019 images in the DUTS-TE dataset, and 4447 images in the HKU-IS dataset. The HKU-IS dataset contains 4447 images.

B. EVALUATION METRICS

In this study, MAE (Mean Absolute Error), F-measure, E-measure, and S-measure were used to evaluate the performance of the algorithm.

The mean absolute error (MAE) was used to calculate the mean absolute difference between the true and predicted graphs, as shown in Eq. (1).

$$MAE = \frac{1}{H \times W} \sum_{i=1}^{W} \sum_{j=1}^{H} |P(i,j) - G(i,j)| \quad (1)$$

where H and W denote the length and width of the image, respectively. where (i, j) denotes the coordinates of the pixel point; $P(i, j)$ denotes the pixel value of the prediction map at (i, j), and $G(i, j)$ denotes the pixel value of the true map at (i, j). Σ denotes the summation and || denotes the absolute value.

The F-measure is the summed average of accuracy and recall, which is a comprehensive evaluation index, as shown in Eq. (2).

$$\begin{cases} F = \dfrac{(1+\beta^2)Precision \times Recall}{\beta^2 \times \Pr ecision + Recall} \\ Precision = \dfrac{TP}{TP+FP} \\ Recall = \dfrac{TP}{TP+FN} \end{cases} \quad (2)$$

where β^2 was typically set to 0.3. Precision is the ratio of the detected significant target pixels to all the predicted significant target pixels. Recall is the ratio of the detected significant target pixels to all the true significant object pixels. TP is the number of pixels predicted to be significant that overlap with the true significant object. FP is the number of pixels that are predicted to be significant, but do not overlap with the true significant object. FN is the number of pixels that are not predicted to be significant, but are in the true significant object. A larger F-measure value indicates better prediction results.

E-measure is a measure of the structural similarity between the predicted significant and true value graphs, as shown in Eq. (3).

$$\begin{cases} \varphi I(i,j) = I(i,j) - \mu I(i,j) \cdot A \\ \xi P(i,j) = \dfrac{2\varphi G(i,j) \circ \varphi P(i,j)}{\varphi G(i,j) \circ \varphi G(i,j) + \varphi P(i,j) \circ \varphi P(i,j)} \\ \phi P(i,j) = f(\xi P(i,j)) \\ f(x) = \dfrac{1}{4}(1+x)^2 \\ E = \dfrac{1}{H \times W} \sum_{i=1}^{W} \sum_{j=1}^{H} \phi P(i,j) \end{cases} \quad (3)$$

where A is a matrix in which all element values are 1 and A has the same size as I, I $\in$ {GT, FM}. φ denotes a bias matrix that calculates the distance between each pixel in a graph and its global mean. $\circ$ denotes the Hadamard product. ξ denotes the alignment matrix, which quantifies the relationship between and by the "convex function" f(x). ϕ denotes an extended alignment matrix that combines pixel-level matching and image-level statistical information through comparison. Larger E-measure values indicated better prediction results.

The S-measure is a metric used to assess the structural similarity between salient and true value graphs. It is derived by measuring the structural similarity between object-aware and region-aware graphs, as shown in Eq. (4).

$$\begin{cases} S = \alpha \times S_o + (1-\alpha) \times S_r \\ S_o = \dfrac{2\mu_F \mu_G + c_1}{\mu_P^2 + \mu_G^2 + c_1} \\ S_r = \dfrac{\mu_{PG} + c_2}{\mu_P + \mu_G + c_2} \end{cases} \quad (4)$$

where P denotes the prediction plot, and G denotes the truth plot. μ_P and μ_G denote the mean of the two and μ_{PG} denotes the structural similarity between P and G. c_1 and c_2 are constants that avoid zero denominators. α is an equalization factor that balances the weights of structural similarity and similarity, typically taking a value of 0.5. S_o is used to measure the similarity between P and G. S_r is used to measure the structure, where a larger S-measure value indicates better prediction.

C. ABLATION EXPERIMENTS

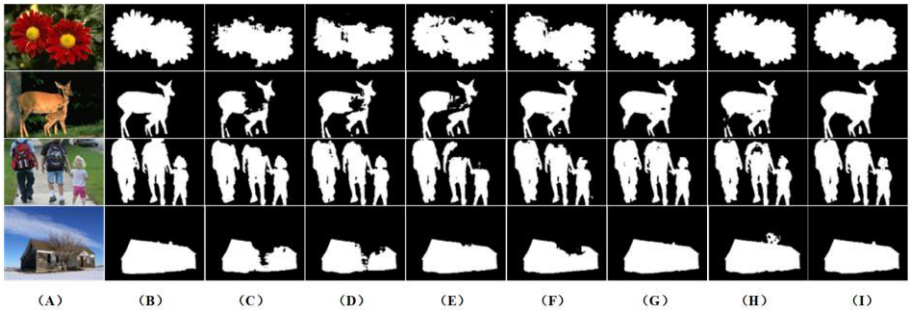
1) ABLATION EXPERIMENTS OF RELATED MODULES

To verify the effectiveness of the methods in this study, different modules and structures were separately added to the base model, and the performance of these modules and structures on the ECSSD dataset was tested. In this study, the strategies in Table 1 were used for optimization, where each row represents the experimental results after combining different methods, networks represent different network structures, base represents the benchmark model, IECA represents the improved version of Improved Efficient Channel Attention (IECA), SAM represents spatial (SAM), MFFM represents the Multi-scale Feature Fusion Module (MFFM), and FEFM represents the Feature Enhancement and Fusion Module (FEFM). Module (FEFM). The ablation experiments were trained on the DUT-OMRONN dataset and tested on the ECSSD dataset, which is less correlated and better reflects the detection performance of the optimized model.

As can be seen from Table 1, the first row represents the results of the benchmark model; the second, third, and fourth rows represent the results of the model after adding only SAM, ECA, IECA, MFFM, and FEFM, respectively, and the detection performance is improved compared to that of the benchmark model; and the fifth row represents the results of the model in this study, and the detection performance is significantly improved. (↓ means a smaller value is better,

TABLE 1. Comparison table of the results of ablation experiments.

Methods	ECSSD			
	MAE↓	F↑	E↑	S↑
Baseline	0.092	0.801	0.830	0.782
Baseline +SAM	0.092	0.803	0.834	0.781
Baseline +ECA	0.078	0.830	0.824	0.801
Baseline +IECA	0.076	0.831	0.849	0.816
Baseline +MFFM	0.065	0.859	0.873	0.842
Baseline +FEFM	0.068	0.858	0.866	0.838
Ours	**0.040**	**0.923**	**0.931**	**0.918**



(A) Original image;(B) ground truth;(C) Baseline;(D) Only add SAM;(E) Only add ECA;(F) Only add IECA;(G) Only add MFFM;(H) Only add FEFM;(I) Ours

FIGURE 6. Comparison of the effect of different modules.

TABLE 2. Comparison of VGG16 and Resnet50 experimental results.

Datasets	VGG16				Resnet50			
	MAE↓	F↑	E↑	S↑	MAE↓	F↑	E↑	S↑
HKU-IS	**0.031**	**0.912**	**0.956**	**0.919**	0.042	0.869	0.92	0.869
ECSSD	**0.040**	**0.923**	**0.931**	**0.928**	0.05	0.876	0.882	0.87
DUT-OMRON	**0.057**	**0.771**	**0.871**	**0.894**	0.076	0.703	0.813	0.755
DUTS-TE	**0.040**	**0.876**	**0.887**	**0.879**	0.058	0.772	0.848	0.816

↑ means a larger value is better. The best results are marked in bold.)

To further evaluate the effectiveness of the SAM, ECA, IECA, MFFM, and FEFM modules, significance plots before and after adding the modules were compared. The results of the ablation experiments are shown in Figure 6.

As shown in Figure 6, the prediction map of the baseline model is not complete and the significance object is incomplete. After adding only SAM, IECA, ECA, MFFM or FEFM modules, the prediction map completeness of the model is greatly improved and the significance object is more complete. It can also be seen that the prediction map of the method in this study is more complete and more closely matches the true value map of the significance object. It further proves that the method in this study can accurately identify the significance objects, which reflects the superiority of the method in this study.

2) ABLATION EXPERIMENTS WITH DIFFERENT BACKBONES

In order to prove the validity of choosing VGG16 as the backbone of this study, training and testing are performed after changing only the backbone while other modules remain unchanged. Four significance detection datasets, ECSSD, DUT-OMRON, DUTS-TE, and HKU-IS, are used as the test set, and the evaluation metrics (MAE, F-value, E-value, and S-value) mentioned above are used, and the evaluation results of VGG16 and Resnet50 are compared, and the evaluation results are shown in Table 2. (↓ means a smaller value is better, ↑ means a larger value is better. The best results are marked in bold.)

Based on the results in Table 2, we can conclude that in the comparison of different datasets, the prediction effect of the model whose backbone network is Resnet50 is worse, while the prediction effect of the model whose backbone network is VGG16 will be better, so in this study, VGG16 is chosen as the backbone network.

D. PERFORMANCE ANALYSIS

Four significance detection datasets, ECSSD, DUT-OMRON, DUTS-TE, and HKU-IS, were used as the test set for the algorithm in this study, using the above evaluation metrics (MAE, F, E, and S). The evaluation results of the proposed algorithm were compared with those of six currently available algorithms: Amulet [22], R3Net [23], PoolNet [24], MINet [16], PurNet [25], and NSAL [26]. The evaluation results are shown in Tables 3 and 4. (↓ indicates that smaller values are better, and ↑ indicates that larger values are better. The best results are marked in bold.)

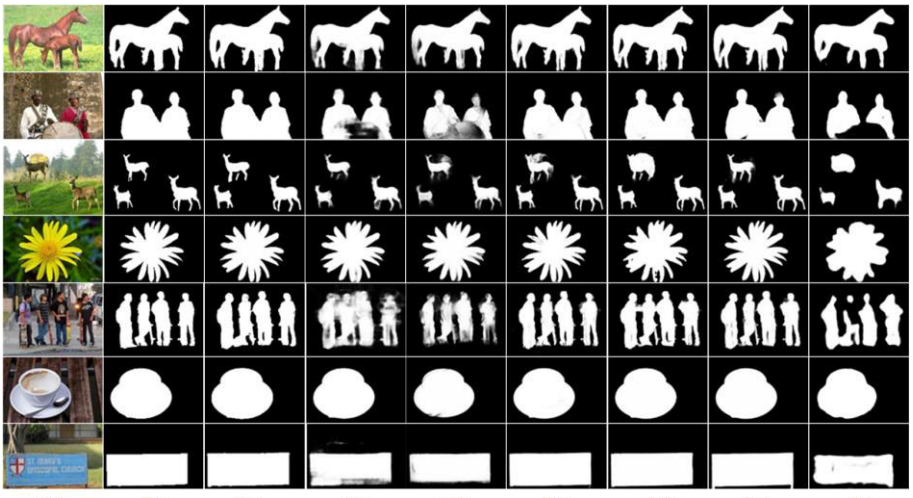
Based on the results in Tables 3 and 4, we can conclude that the proposed method performed well in detecting significant objects and outperformed the other models on most of the datasets. When comparing different datasets, the algorithm in this study performed the best on the HKU-IS dataset. Compared with the Amulet algorithm, the MAE of the algorithm in this study decreased by 0.021, indicating that the objective localization accuracy of the algorithm in this study was higher. Compared with the NSAL algorithm, the F of the algorithm in this study improved by 0.048, which is a better comprehensive performance, E improved by 0.033, and S improved by 0.065, which indicates that the prediction map generated by the algorithm in this study has a higher structural similarity with the original true value map. This proves that the proposed algorithm exhibits good robustness and accuracy in a boundary fuzzy scenario. On the

TABLE 3. Comparison of MAE, F, E and S metrics for different algorithms on the HKU-IS and ECCSD.

	HKU-IS				ECSSD			
Methods	MAE ↓	F ↑	E ↑	S ↑	MAE ↓	F ↑	E ↑	S ↑
Amulet2017	0.052	0.889	-	-	0.062	0.911	-	0.894
R3Net2018	0.038	0.893	0.939	0.895	0.046	0.914	0.929	0.910
PoolNet2019	0.032	0.899	0.949	0.916	0.039	0.915	0.924	0.921
MINet2020	**0.031**	0.904	0.948	0.912	**0.036**	0.922	0.923	0.919
PurNet2021	0.036	0.894	**0.956**	0.918	0.040	0.92	**0.953**	0.925
NSAL2022	0.051	0.864	0.923	0.854	0.077	0.856	0.884	0.834
Ours	**0.031**	**0.912**	**0.956**	**0.919**	0.040	**0.923**	0.931	**0.928**

TABLE 4. Comparison of MAE, F, E and S metrics of different algorithms on DUT-OMRON and DUTS-TE.

	DUT-OMRON				DUTS-TE			
Methods	MAE ↓	F ↑	E ↑	S ↑	MAE ↓	F ↑	E ↑	S ↑
Amulet2017	0.098	0.737	-	0.780	0.075	0.773	-	0.803
R3Net2018	0.061	**0.792**	**0.939**	0.817	0.059	0.785	0.867	0.834
PoolNet2019	0.056	0.786	0.869	0.836	0.040	0.809	0.889	0.883
MINet2020	0.057	0.741	0.857	0.822	**0.039**	0.823	0.895	**0.884**
PurNet2021	**0.054**	0.768	0.876	0.841	0.043	0.816	**0.915**	0.871
NSAL2022	0.088	0.648	0.801	0.745	0.073	0.73	0.849	0.781
Ours	0.057	0.771	0.871	**0.894**	0.040	**0.876**	0.887	0.879



(A) Original image; (B) ground truth;(C) Ours; (D) AmulNet; (E) R3Net; (F) PoolNet; (G) MINet; (H) PurNet; (I) NSAL

FIGURE 7. Visualization comparison between this method and other 6 methods.

DUT-OMRON dataset, the MAE of the algorithm in this study decreased by 0.041, F improved by 0.123, E improved by 0.07, and S improved by 0.149 compared with the Amulet algorithm, which proves that the algorithm in this study has a good detection performance in complex scenes. On the ECSSD dataset, this algorithm ranks first in F and S, and although MAE and E are not first, they are only 0.004 and 0.022 lower than the first place, respectively, which shows that this algorithm has a good generalization performance. For the DUTS-TE dataset, the MAE, F, and S of the algorithm in this study were in the top three, and only E was slightly lower than those of the other algorithms. This proves that the method used in this study has a high accuracy and robustness in a single scenario. By combining the evaluation index data

of the four datasets, it can be concluded that the proposed algorithm has the advantages of high localization accuracy, good comprehensive performance, and high structural similarity for significant object detection.

To further evaluate the effectiveness of the proposed method, the significance plots of this study's model method and those of the other six methods were compared, and the results are shown in Figure 7. Figure 7 shows the following situations: (1) the scene contains multiple saliency targets, (2) the background is complex, and (3) the contrast between the foreground and background is low.

It is clear from Figure 7 that the existing advanced methods suffer from blurred detection of the target edges (methods D, E, and H), poor detection integrity (methods G and I), and foreground and background misclassification (methods F and I). In contrast, it is further demonstrated that the method proposed in this study can accurately identify all regions of salient targets, reflecting the good generalization performance, accuracy, and robustness of the method in this study.

IV. CONCLUSION

SOD has been widely used in scenarios, such as robot navigation, semantic segmentation, object recognition, and detection. Currently, SOD models suffer from many problems, such as front and back view misclassifications and edge blur. Based on the existing research, this study proposes a multi-scale feature enhancement method for a saliency object detection algorithm. The proposed algorithm mainly uses the multi-scale module and the feature enhancement module to increase the network's attention to salient objects. It also makes use of the attention module to reduce the interference from the background. Thus, the overall performance of the network is improved. From the comparison results between the proposed algorithm and the other six algorithms, it can be known that the proposed algorithm achieves better detection results. From the visualized comparison graph, it can be seen that the proposed algorithm is able to accurately identify all regions of the salient object. Although the proposed algorithm has achieved better results, there is still room for progress in the part of local and global information fusion, and the next step will be to conduct in-depth research in this area.

DATA AVAILABILITY

Data used to support the findings of this study are available from the corresponding author upon request.

CONFLICTS OF INTEREST

There is no conflict of interest regarding the publication of this study.

REFERENCES

[1] Z. Q. Wang, Y. S. Zhang, Y. Yu, J. Min, and H. Tian, "Review of deep learning based salient object detection," *J. Image Graph.*, vol. 2022, no. 7, pp. 2112–2128, 2022.

[2] H. B. Bi, H.-H. Zhu, L. N. Yang, C. Zhang, and R. W. Wu, "Design of video salient object detection system based on multi-level feature fusion," *Res. Explor. Lab.*, vol. 2022, no. 3, pp. 94–98, 2022.

[3] X. W. Chen, Y. Zhang, J. J. Lin, and Q. Zhang, "Global context guided multi-scale feature network for salient objectdetection," *Comput. Appl. Softw.*, vol. 2022, no. 3, pp. 146–153, 2022.

[4] W. R. Li, D. Xu, J. L. Shi, and S. C. Huang, "Review of salient object detection: Models, applications and prospects," *Appl. Res. Comput.*, vol. 2022, no. 7, pp. 1941–1950, 2022.

[5] X. F. Zhou, S. Y. Guo, H. F. Wen, B. T. Liu, S. F. Li, J. Y. Zhang, and C. G. Yan, "Deep learning-based co-salient object detection on RGBD images," *J. Signal Process.*, vol. 2022, no. 6, pp. 1213–1221, 2022.

[6] M. Guo, Y. Zhao, C. Zhang, and Z. Chen, "Fast object detection based on selective visual attention," *Neurocomputing*, vol. 144, pp. 184–197, Nov. 2014.

[7] S. Wang, M. Wang, S. Yang, and L. Jiao, "New hierarchical saliency filtering for fast ship detection in high-resolution SAR images," *IEEE Trans. Geosci. Remote Sens.*, vol. 55, no. 1, pp. 351–362, Jan. 2017.

[8] X. Qin, S. He, X. Yang, M. Dehghan, Q. Qin, and J. Martin, "Accurate outline extraction of individual building from very high-resolution optical images," *IEEE Geosci. Remote Sens. Lett.*, vol. 15, no. 11, pp. 1775–1779, Nov. 2018.

[9] Y. Wang, R. Wang, X. Fan, T. Wang, and X. He, "Pixels, regions, and objects: Multiple enhancement for salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 10031–10040.

[10] H. Zhou, B. Qiao, L. Yang, J. Lai, and X. Xie, "Texture-guided saliency distilling for unsupervised salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 7257–7267.

[11] X. Tian, J. Zhang, M. Xiang, and Y. Dai, "Modeling the distributional uncertainty for salient object detection models," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2023, pp. 19660–19670.

[12] W. Zhang, L. Zheng, H. Wang, X. Wu, and X. Li, "Saliency hierarchy modeling via generative kernels for salient object detection," in *Proc. Eur. Conf. Comput. Vis.* Cham, Switzerland: Springer, 2022, pp. 570–587.

[13] L. Q. Cui, J.-J. Chen, Q. Y. Ren, and B. H. Wang, "Saliency object detection based on multiple features and prior information," *J. Image Graph.*, vol. 2020, no. 2, pp. 321–332, 2020.

[14] Y. B. Zhang and F. Zhang, "Salient object detection based on texture and color features," *Comput. Digit. Eng.*, vol. 49, no. 9, pp. 1793–1798 and 1877, 2021.

[15] Z. Ouyang, H. Zhang, Z. Tang, L. Peng, and S. Yu, "Research on defect detection algorithm of complex texture ceramic tiles based on visual attention mechanism," *Xibei Gongye Daxue Xuebao/J. Northwestern Polytech. Univ.*, vol. 40, no. 2, pp. 414–421, Apr. 2022.

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

[17] J. S. Fang, Y. H. Tao, G. P. Zhu, and Y. Y. Chen, "Saliency detection from complex background with attention-based boundary-aware pyramid pooling network," *Comput. Eng. Appl.*, vol. 2023, no. 4, pp. 1–11, 2023.

[18] X. J. Wang, M. Y. Li, L. Wang, F. Liu, and W. Wang, "Salient object detection method based on multi-scale feature-fusion guided by edge information," *Infr. Laser Eng.*, vol. 52, no. 1, pp. 261–270, 2023.

[19] X. Yang, H. L. Zhu, and G. J. Mao, "Dual-stream fusion and edge-aware network for salient object detection," *Comput. Eng. Appl.*, vol. 2023, no. 4, pp. 1–12, 2023.

[20] Q. Wang, B. Wu, P. Zhu, P. Li, W. Zuo, and Q. Hu, "ECA-Net: Efficient channel attention for deep convolutional neural networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2020, pp. 11531–11539.

[21] S. Woo, J. Park, J.-Y. Lee, and I. S. Kweon, "CBAM: Convolutional block attention module," 2018, *arXiv:1807.06521*.

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

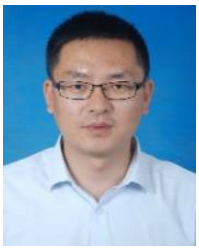
[23] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, and P.-A. Heng, "R3net: Recurrent residual refinement network for saliency detection," in *Proc. 27th Int. Joint Conf. Artif. Intell.*, Palo Alto, CA, USA, Jul. 2018, pp. 684–690.

[24] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3912–3921.

[25] J. Li, J. Su, C. Xia, M. Ma, and Y. Tian, "Salient object detection with purificatory mechanism and structural similarity loss," *IEEE Trans. Image Process.*, vol. 30, pp. 6855–6868, 2021.
[26] Y. Piao, W. Wu, M. Zhang, Y. Jiang, and H. Lu, "Noise-sensitive adversarial learning for weakly supervised salient object detection," *IEEE Trans. Multimedia*, vol. 25, pp. 2888–2897, 2023.

SU LI received the B.S. degree from the Yancheng Institute of Technology, Yancheng, China, in 2021, where she is currently pursuing the M.Eng. degree. Her current research interests include computer vision technology and intelligent control systems and signal detection.

RUGANG WANG received the B.S. degree from the Wuhan University of Technology, Wuhan, China, in 1999, the M.S. degree from Jinan University, Guangzhou, China, in 2007, and the Ph.D. degree from Nanjing University, Nanjing, China, in 2012. Currently, he is a Professor with the College of Information Engineering, Yancheng Institute of Technology, Yancheng, China. His current research interests include optical communication networks, novel and key devices for optical communication systems, and image processing technology.



FENG ZHOU received the B.S. and M.S. degrees from Southeast University, Nanjing, China, in 2004 and 2012, respectively. He is currently pursuing the Ph.D. degree with the Army Engineering University of PLA. Since 2017, he has been an Associate Professor with the College of Information Engineering, Yancheng Institute of Technology, Yancheng, China. His current research interests include cooperative communication, computer vision technology, and image processing technology.

YUANYUAN WANG, photograph and biography not available at the time of publication.

NAIHONG GUO, photograph and biography not available at the time of publication.