DOI: https://doi.org/10.61187/ita.v1i1.14

Salient object detection algorithm based on diversity features and global guidance information

Fan Shao, Kai Wang, Yanluo Liu*

Shaanxi University of Science and Technology, School of Electronic Information and Artificial Intelligence, Xi 'an, Shaanxi 710021
Correspondence Author[*]

Abstract: ***Aiming at the problems of traditional salient object detection methods such as fuzzy boundary and insufficient information integrity, a salient object detection network composed of feature diversity enhancement module, global information guidance module and feature fusion module is proposed. Firstly, asymmetric convolution, cavity convolution and common convolution are spliced to form a feature diversity enhancement module to extract different types of spatial features corresponding to each feature layer. Secondly, the global information guidance module transmits the information captured by the coordinate attention mechanism to each feature layer through the global guidance stream, so as to learn the semantic relationship between different feature layers and alleviate the dilution effect; Finally, the feature fusion module receives the high-level features output from the previous layer, the low-level features of the corresponding layer and the global context information generated by the global information guidance module, and the cascade feature diversity enhancement module gradually integrates the multi-level features to refine the saliency feature map. Comparative experiments on five public data sets show that this method can achieve the highest values of 0.959 and 0.030 in F-measure and MAE. Compared with other seven advanced algorithms, it has better detection performance.***

Keywords: Salient object detection; Global information guidance; Diversity character; Feature fusion**.**

1. Introduction

Saliency target detection automatically perceives salient targets and aims to quickly locate and segment the most salient objects for human vision, playing a crucial role in downstream tasks such as image editing[1] and visual tracking[2]. Early salient target detection generally relied on low-level features such as color, texture, contrast, and orientation or used saliency priors such as center prior and boundary prior for heuristically combining low-level features to improve saliency estimation. For example, Chenget al[3] focused on the fact that the saliency of a region depends primarily on its contrast relative to nearby regions, and defined the saliency value of individual pixel points as the contrast between that pixel and other pixels in the image by evaluating global contrast differences and spatially weighted consistency scores. Liu et al[4] hypothesized that the saliency of each pixel point in an image is related to the most representative salient element, integrating learning strategies into visual partial differential equation techniques to transform bottom-up and top-down information into visual attention in saliency diffusion through an adaptive system of learning partial differential equations. However, due to the lack of high-level semantic information, the salient object detection may fail when the salient object deviates from the image center and is close to the image boundary or has obvious overlap with the image boundary.

With the development of deep learning techniques, fully convolutional neural networks have shown strong feature extraction capabilities and are gradually being widely used in the field of saliency target detection[5].For example, Qin et al[6] iteratively predicted saliency maps or intermediate feature maps at multiple scales by learning the residuals between features at each level in order to capture rich spatial and multi-scale information, and improved the boundary quality using a hybrid loss function.Wu et al[7] used the generated saliency feature maps to recursively optimize deep features and cascaded multiple modules for accurate detection of salient targets. Wang et al[8] used attention mechanism for saliency enhancement of extracted features, preserving edge details as well as enhancing semantic information by making full use of low-level features, and refining edge details of salient objects using aggregation modules. Chen et al[9] extracted and enhanced features by spatial and channel attention modules, and used global contextual information flow to mitigate the dilution problem of feature delivery process. Although the above multiscale fusion networks exploit the complementarity of high-level and low-level features to enhance the performance of network models by fusing multi-layer features at different scales as much as possible. However, extracting feature information by layer-by-layer abstraction will face the problem of low resolution of high level features and high resolution of low level features, and only local features can be observed or too many invalid features are obtained.

To solve the above problem, multi-branch networks are usually designed as a network structure with multiple outputs, and the task of extracting saliency features is performed side by side with other subtasks, and the subtasks are made to gain to the saliency feature extraction task by using joint training, etc. For example, Zhou et al[10] constructed a two-branch decoder for obtaining image saliency and contour information, explored the correlation between saliency and contour by transmitting features interactively, and used an adaptive contour loss function to improve the performance of saliency target detection. Liu et al[11] used an encoder-decoder structure to connect the features of adjacent layers to capture saliency targets at different scales in the image, and used a parallel multibranch structure to achieve feature extraction under different perceptual fields.Liu et al[12] designed two modules based on pooling techniques to provide location information of saliency targets and refine salient object features for different layers, respectively, using boundary information and saliency information for joint training.Qin et al[13] did not use any pre-trained model and trained a two-layer nested U-shaped structure from scratch, using pooling techniques and residual blocks to capture the contextual information. In addition, some researchers have introduced capsule networks[14] into the saliency target detection task to further improve the completeness of salient objects. Although the above-mentioned studies are continuously improving the performance of saliency detection models, how to make full use of contextual and boundary information and designing better network structures are still key factors in extracting saliency regions. For example, when the features of the background noise are similar to those of the salient target, the network model is very likely to produce wrong or missed detection during feature extraction; when the shape and contour of the salient target are complex, the network model will be difficult to accurately define the boundary location of the salient target.

In this paper, to address the problem of missing information of salient regions, we design a feature diversity enhancement module to mine three different types of spatial features of each feature layer, and use a coordinate attention mechanism to further focus the location of salient targets, use global contextual information flow to avoid feature dilution, and use a feature fusion module to aggregate feature information of different levels. To address the problem of blurred boundaries, a deep supervision strategy and a hybrid loss function are used to promote the precise definition of target boundaries in the network model, and the cascaded feature diversity enhancement module and the feature fusion module are used to gradually refine the saliency features.

2. Network Structure

As shown in Fig.1, this paper is based on the encoder decoder structure, and the encoder selects the pre-trained ResNet50[15] as the backbone network, C1~C5 represents the corresponding five residual blocks. Due to the large size of the feature space extracted by C1, the encoder uses only the features of the C1~C5, in order to improve the computational efficiency. First, the low-resolution features of the highest layer are sent to the feature diversity enhancement module to generate feature maps with features extracted by null convolution, asymmetric convolution, and ordinary convolution; then, the coordinate attention module is used to generate global semantic information to guide the decoding process; finally, the features of the previous layer that have passed through the feature diversity enhancement module are sent to the feature fusion module together with the features extracted by the backbone network at the current layer and the global semantic information Finally, the features extracted from the upper layer of the feature diversity enhancement module and the current layer of the backbone network together with the global semantic information are sent to the feature fusion module for effective feature aggregation. The cascaded feature diversity enhancement and feature fusion modules gradually refine the saliency feature maps under the supervision of the hybrid loss function. Each module is described separately in the next section.

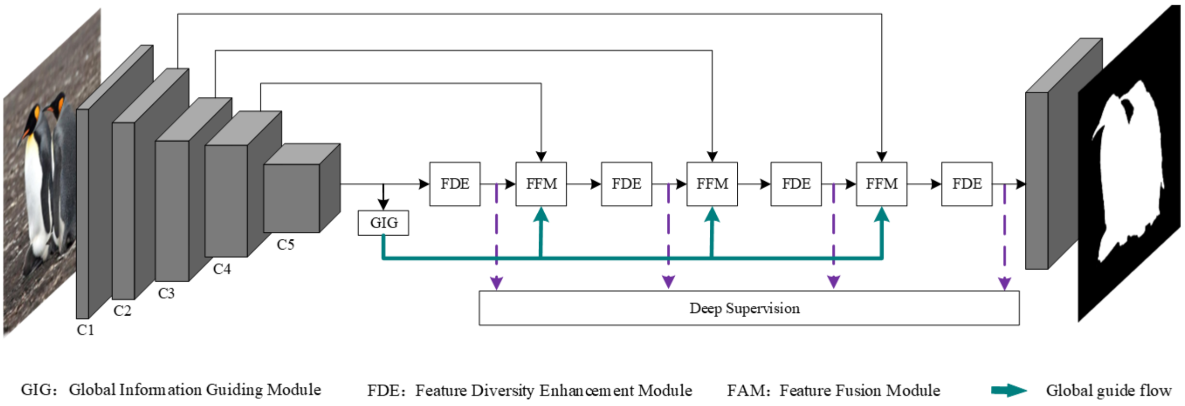
C1
C2
C3
C4
C5
GIG
FDE
FFM
FDE
FFM
FDE
FFM
FDE
Deep Supervision
GIG：Global Information Guiding Module
FDE：Feature Diversity Enhancement Module
FAM：Feature Fusion Module
Global guide flow

Fig.1 Network Structure

2.1 Feature Diversity Enhancement Module, FDE

The location of salient targets varies greatly from image to image, and designing the appropriate size and shape of the convolution kernel plays a key role in improving the feature learning ability of the network. Since larger convolutional kernels are easier to extract global information of images and smaller convolutional kernels are easier to extract local information of images, most methods use simple stacking of convolutional layers of different sizes to extract features. In this paper, we use three different shapes of convolutional stitching of ordinary convolution, asymmetric convolution[16] and atrous convolution [17] to form a feature diversity enhancement module, thus generating a feature map that incorporates three spatial features to further enrich the spatial feature information. In addition, the feature diversity enhancement module is used separately for feature maps of different resolutions to enhance the completeness of significant information extraction.

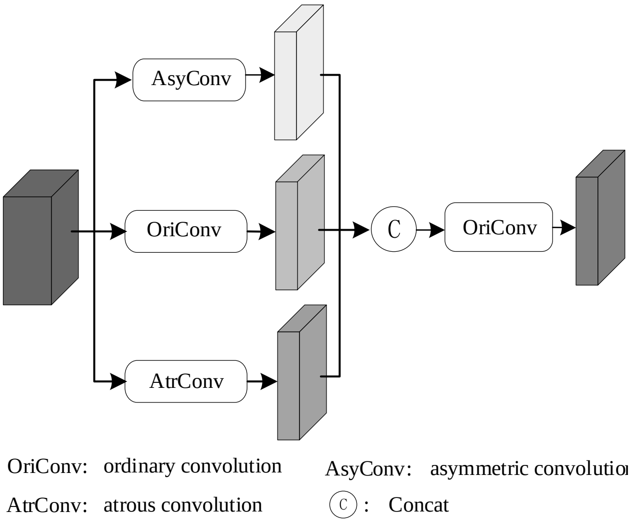
AsyConv
OriConv
AtrConv
C
OriConv
OriConv: ordinary convolution
AsyConv: asymmetric convolutio
AtrConv: atrous convolution
C : Concat

Fig.2 Feature Diversity Enhancement Module

As shown in Fig.2, where the ordinary convolution uses a 3×3 convolution kernel; the asymmetric convolution replaces direct 3×3 convolution by 3×1 convolution and 1×3 convolution; and the hole convolution sets the expansion rate to 2 to expand the perceptual field by adding holes while avoiding loss of resolution by downsampling. The specific operation is expressed as follows:

$$F(i) = f_{ori}(f_{ori}(f(i)) \oplus f_{asy}(f(i)) \oplus f_{atr}(f(i))) \qquad (1)$$

Where, $f(i)$ denotes the four feature layers in the cascade structure, and when $i=1$, $f(1)$ denotes the low resolution extracted from the highest layer of the backbone network. When $2 \le i \le 4$, $f(i)$ indicates the refined feature map output by the cascaded feature fusion module. $F(i)$ denotes the feature extraction result of the feature diversity enhancement module, $f_{ori}(i)$ represents the normal convolution, $f_{asy}(i)$ represents the asymmetric convolution, $f_{atr}(i)$ represents the null convolution, and $\oplus$ represents the splicing operation. Finally, $F(i)$ the batch normalization and ReLU activation function operations are performed to further enrich the diversity of the extracted features in each feature layer and capture the salient objects of different sizes and salient features at different locations.

2.2 Global Information Guiding Module, GIG

Deep learning-based convolutional neural networks often suffer from feature dilution during top-down information transfer, producing salient graphs with incomplete targets or unclear boundaries. In this paper, a global guidance information module consisting of a coordinate attention module[18] and a series of global guidance streams is constructed. The coordinate attention module, compared with other attention modules, considers not only the relationship between channels but also the location information in feature space, which is conducive to better localization and identification of saliency regions. As shown in Fig. 3, the global information guidance module is connected to the top-level features *C5* of the backbone network and adjusts the feature weights of different regions in the network by averaging pooling and 1×1 convolution, and generates *C5* a pair of direction-aware and position-sensitive codes by combining channel and direction information, which are applied to the input feature map *C5* in a complementary way by multiplying pixel points to enhance the attention to salient objects. The specific procedure is as follows:

First, the global average pooling operation is used to decompose the feature map *C5* into feature codes along the vertical *X* and horizontal *Y* directions to obtain two independent direction-aware feature maps f_x and f_y in the vertical and horizontal directions, each of which captures the long-range dependence of the spatial direction and preserves the accurate location information. Denote the feature map size as $W \times H$, $C5(i, j)$ denote the value of the feature map *C5* at the position as (i, j) expressed by the equation:

$$f_x = \frac{1}{W} \sum_{0 \le i \le W}^{i} C5(x,i) \qquad (2)$$

$$f_y = \frac{1}{H} \sum_{0 \le j \le H}^{j} C5(y,j) \qquad (3)$$

Secondly, the feature maps f_x and f_y cascade embedded with specific direction information and used to capture the relationship between channels by 1×1 convolution blocks, and the pixel points of different channels are linearly combined by convolution operations, and then nonlinear operations such as h-swish are performed to make full use of the captured position information so that the significant target regions can be captured accurately. Let f_{conv} denote the convolution operation and $\oplus$ denote the cascade operation, then the equation is expressed as:

$$F = f_{conv}(f_x \oplus f_y) \qquad (4)$$

Finally, the attentional feature map with channel and spatial features F is cut into g_x and g_y, and the weights of g_x and g_y are adjusted again by 1×1 convolution and Sigmoid activation function, at which time, g_x and g_y are a pair of feature maps with direction-aware and location information, and the input feature maps *C5* and g_x and g_y are multiplied by pixel points to encode the channel relationship and long-range dependence of the image by the captured accurate location information to realize the characterization of saliency regions. The formula is expressed as follows:

$$g_x, g_y = f_{split}(F) \qquad (5)$$

$$G(k) = C5 * f_{conv}(g_x) * f_{conv}(g_y) \qquad (6)$$

Where, f_{split} denotes the feature cut function and $G(k)$ denotes the global information guide flow. In order to pass the salient target features extracted by the above method to the cascaded feature layers with different resolutions, this paper designs a series of global bootstrap flows in the global information guidance module, and makes the location of the salient object explicitly known in each layer of the feature map through the global guidance information weights during the top-down passing process, thus ensuring that the localization information is not diluted.

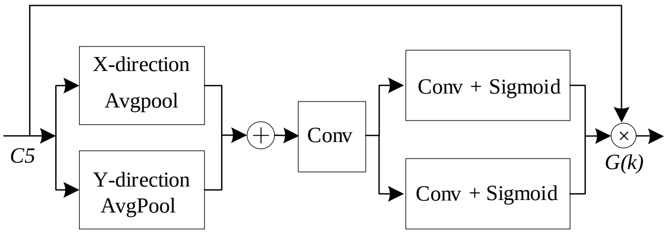

Fig.3 Global Information Guiding Module

2.3 Feature Fusion Module, FFM

In the decoding process, how to reasonably fuse the global guidance flow of the above global information guidance module with the feature maps of different layers, and try to avoid the noise of the low-level feature maps and the impact of the fuzzy boundaries of the high-level feature maps on the model performance becomes an urgent problem to be solved. The feature fusion module used in this paper contains 4 sub-branches to receive high-level features $F(i+1)$ from the previous layer, as shown in Fig.4. First, the forward-passed feature maps obtained from the feature enhancement module are downsampled using an average pooling operation with different downsampling rates to obtain different sizes of perceptual fields, where the sampling rates are set to 2, 4, and 8, Then, the feature maps of each branch are convolved and upsampled to the same size as the input features, and the up-sampled feature maps of the four branches are fused together by the addition operation to reduce the jagged effect caused by upsampling by the convolution operation.

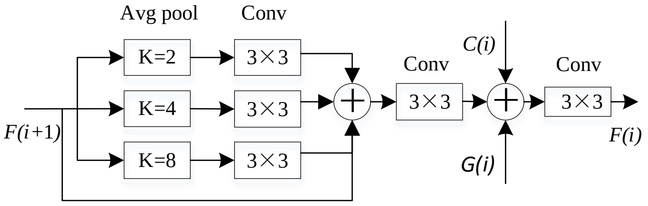

Fig.4 Feature Fusion Module

2.4 Loss function

Inspired by the boundary-aware hybrid loss for saliency target detection proposed by Qin et al[7], the hybrid loss function in this paper consists of Binary CrossEntropy Loss (BCE Loss)[19], Multiscale Structure Similarity Loss (MS-SSIM Loss)[20], and Intersection over Union Loss (IoU Loss)[21], which to achieve different levels of saliency information extraction at pixel level, block level, and image level. The formula is expressed as:

$$L_{hybrid} = L_{bce} + L_{ms-ssim} + L_{iou} \tag{7}$$

Binary cross-entropy is a loss function widely used in significance target detection tasks to measure the degree of variation of different probability distributions of the same random variable in a binary classification model, and is calculated as follows:

$$L_{bce} = -\sum_{i=1}^{N} y_i \ln(\sigma(x_i)) - \sum_{i=1}^{N} (1-y_i)\ln(1-\sigma(x_i)) \tag{8}$$

where, N denotes the number of samples, $x_i \in \{0,1\}$ denotes the labels of the samples i, and y_i denotes the predicted values. Since the binary cross-entropy only considers the loss value of each pixel, and the structural similarity loss function measures the degree of distortion as well as the degree of similarity of the image from three key features: brightness, contrast and structure, this paper introduces MS-SSIM Loss into the training process so as to learn the global structural information of the image. The formula is expressed as:

$$L_{ms-ssim} = 1 - \prod_{m=1}^{M} \left(\frac{2\mu_p\mu_g + c_1}{\mu_p^2 + \mu_g^2 + c_1}\right)^{\beta_m} \times \prod_{m=1}^{M} \left(\frac{2\sigma_{pg} + c_2}{\sigma_p^2 + \sigma_g^2 + c_2}\right)^{\gamma_m} \tag{9}$$

where, M denotes the total number of scales, μ_P, μ_g and σ_p, σ_g are the means and variances of p, g, respectively, σ_{pg} and denotes their covariances. $c_1 = 0.001^2$ and $c_2 = 0.003^2$ are to avoid the case of zero denominator. β_m, γ_m is used to define the relative importance.

The cross-ratio loss function considers the pixel values of each region, and it is mainly used to measure the overlap rate between the predicted results and the true labels, independent of the scale variation and distribution imbalance. The formula for IoU Loss is expressed as:

$$L_{iou} = 1 - \sum_{r=1}^{H}\sum_{c=1}^{W} S(r,c)G(r,c) \times \frac{1}{\sum_{r=1}^{H}\sum_{c=1}^{W}[S(r,c)+G(r,c)-S(r,c)G(r,c)]} \tag{10}$$

Where, H, W denote the height and width, $G(r,c) \in \{0,1\}$ denote the true value at (r,c) the location, and $S(r,c)$ denote the predicted value at the location (r,c). In addition, a deep supervision strategy is used during the training process to improve the learning ability among the layers, and a mixed loss function is used for iterative training of the model. Since the model has four side outputs, the four loss functions of the four layers and the sum of the loss functions are calculated separately to obtain a 5 × 4 array, and by continuously updating the network parameters, the smaller the loss value is, the better the obtained network model parameters are until convergence.

3. Experiments and results analysis

3.1 Data set and experimental setup

To evaluate the effectiveness of the proposed method, five representative and commonly used saliency target detection datasets were selected for model training and testing, including: the DUTS[22], PASCAL-S[23], ECSSD[24], SOD[25], and DUT-OMRON[26] datasets. Among them, DUTS is a benchmark dataset containing DUTS-TR and DUTS-TE with 10,553 training images and DUTS-TE with 5,019 test images, which is the largest saliency target detection dataset to date. ECSSD contains 1,000 images of complex scenes with large size differences. PASCAL-S contains SOD contains 300 images of more complex scenes with multiple salient targets. dut-omron contains 5168 high-quality images, each containing at least one or two structurally complex foreground objects, and is the most challenging salient target detection dataset to date.

In this paper, DUTS-TR in DUTS dataset is used as the model training dataset, and other datasets are used for algorithm effect testing. The network structure is based on the Pytorch deep learning framework, using ResNet50 as the backbone network to extract the basic features of the image, and loading the pre-trained weights from ImageNet as the initial weights of the network, with the model trained for a total of The model is trained for a total of 26 rounds (epoch), the batch size is set to 16, the Adam optimizer is used, the initial learning rate is set to 5e-5, the weight decay is set to 5e-4, and after 15 rounds of training, the learning rate is reduced to 5e-6. The data enhancement method uses horizontal flipping, and the image size is uniformly adjusted to 320×320 for training.

3.2 Evaluation indicators

The accuracy-recall curve (P-R curve), Mean Absolute Error (MAE), and F-measure (F) are commonly used as performance evaluation metrics for significance target detection. Among them, the P-R curve compares the binarized prediction graph with the true value graph, and after calculating the accuracy and recall, the curve can be plotted with the accuracy as the vertical coordinate and the recall as the horizontal coordinate. The calculation formula is as follows:

$$precision = \frac{TP}{TP+FP} \text{,} \quad recall = \frac{TP}{TP+FN} \qquad (11)$$

where, represents true positives, represents false positives, and represents false negatives. represents the accuracy rate, which is the percentage of true positives among predicted positives from the prediction perspective. represents the recall rate, which is the percentage of predicted positives among true positives from the actual perspective.

The F-measure is a weighted average of the calculated accuracy and recall, and is a comprehensive evaluation metric. It is usually set to 0.3 and the formula is expressed as follows:

$$F_{\beta} = \frac{(1+\beta^2)\, precision \times recall}{\beta^2\, precision + recall} \qquad (12)$$

The mean absolute error is used to evaluate the performance of the algorithm directly by calculating the error between the predicted and true plots in combination with the significance target and the difference between the background and true plots.

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} | P(x,y) - G(x,y) | \qquad (13)$$

where, refers to the prediction map, represents the true value, and and represents the width and height, respectively. Specifically, the larger the area enclosed by the curve and the coordinate axis, the better the performance of the algorithm, the larger the F-measure value, the higher the accuracy of the predicted significance image, and the smaller the mean absolute error, the better the performance.

3.3 Comparison with other advanced methods

To ensure the persuasiveness of the experiments and the fairness of the experiments, comparisons were made with seven current mainstream saliency target detection algorithms, including PFSNet[5], BASNET[6], CPD[7], GCPANet[9], ITSD[10], PoolNet[12], and U2Net[13]. All the prediction maps used for the evaluation used the original significant maps provided by the officials, and all methods were evaluated using the same evaluation code.

Tab.1 shows the metrics evaluation results of the model in the paper and the other seven advanced methods on five datasets. As can be seen, compared with the other seven advanced methods, the proposed model in the paper obtains optimal or suboptimal results on the ECSSD, DUTS-TE, PASCAL-S, SOD, and DUT-OMRON datasets, and the optimal results are highlighted in bold font, where F denotes F-measure, and the larger the value is the better the algorithm performance, and the smaller the value of MAE indicates the algorithm performance of the algorithm is better. Among them, on the ECSSD dataset, the F-measure and MAE metrics of the model proposed in the paper were tested at 0.959 and 0.030, respectively, which showed a 1.7% improvement in F-measure and a 0.7% reduction in MAE compared to BASNet, the boundary-aware network proposed by Qin[6] et al. On the most challenging DUT-OMRON dataset The test results show that F-measure metrics improve by 1.5% and MAE decreases by 0.2% compared to BASNet. Therefore, it can be concluded that although the method in this paper uses a similar hybrid loss function and model training method as BASNet, it has a better detection effect, which further reflects that the network model consisting of feature diversity enhancement module, global information guidance module, and feature fusion module constructed in this paper can better capture the semantic features of salient objects and narrow the gap between the saliency prediction map and the true label values.

Tab.1 Quantitative comparison

METHOD	ECSSD		DUTS-TE		PASCAL		SOD		DUT-OMRON	
	F↑	MAE↓	F↑	MAE↓	F↑	MAE↓	F↑	MAE↓	F↑	MAE↓
PFSNet[5]	0.952	0.032	**0.898**	0.036	**0.881**	0.063	——	——	**0.823**	0.055
BASNet[6]	0.942	0.037	0.860	0.048	0.857	0.075	0.849	0.112	0.805	0.056
CPD[7]	0.939	0.037	0.865	0.043	0.864	0.072	0.857	0.110	0.796	0.055
GCPANet[9]	0.948	0.035	0.888	0.038	0.876	**0.060**	0.872	**0.087**	0.811	0.056
ITSD[10]	0.947	0.035	0.882	0.041	0.877	0.071	0.880	0.095	0.820	0.060
PoolNet[12]	0.945	0.039	0.880	0.040	0.869	0.074	0.867	0.100	——	——
U2Net[13]	0.950	0.033	0.872	0.045	0.862	0.076	0.861	0.108	0.822	**0.054**
Ours	**0.959**	**0.030**	**0.898**	**0.033**	0.880	**0.060**	**0.884**	0.091	0.820	**0.054**

The P-R curves are plotted according to the test results, as shown in Fig.5. Among them, the thick red solid line indicates the test results of the model proposed in the paper, and it can be seen that the P-R curve of the algorithm in this paper has the largest area enclosed by the coordinate axes, which mainly benefits from the rich spatial feature information extracted by the feature diversity enhancement module mentioned in the paper, and the global information guidance module uses the coordinate attention mechanism to focus on the salient target area, and further transmits the location of the salient target through the global guidance flow to each feature layer, which enhances the ability of locating the salient target in complex scenes. The global information guidance module uses the coordinate attention mechanism to focus on the salient target area, and further transmits the location information of the salient target to each feature layer through the global guidance flow, which enhances the ability of locating the salient target in complex scenes. At the same time, the feature fusion module plays the role of rational use of the input feature information and noise suppression. Similarly, F-measure is the weighted average of accuracy and recall, and when there is a conflict between accuracy and recall, the comprehensive evaluation index F-measure can be used to reflect the overall situation. From the plotted results, it can be intuitively seen that the test results obtained using the method in the paper under the same data set plot the F-measure curve at the top of all the compared algorithms, meaning that the area enclosed with the coordinate axis is the largest, which once again verifies the effectiveness of the method proposed in the paper.

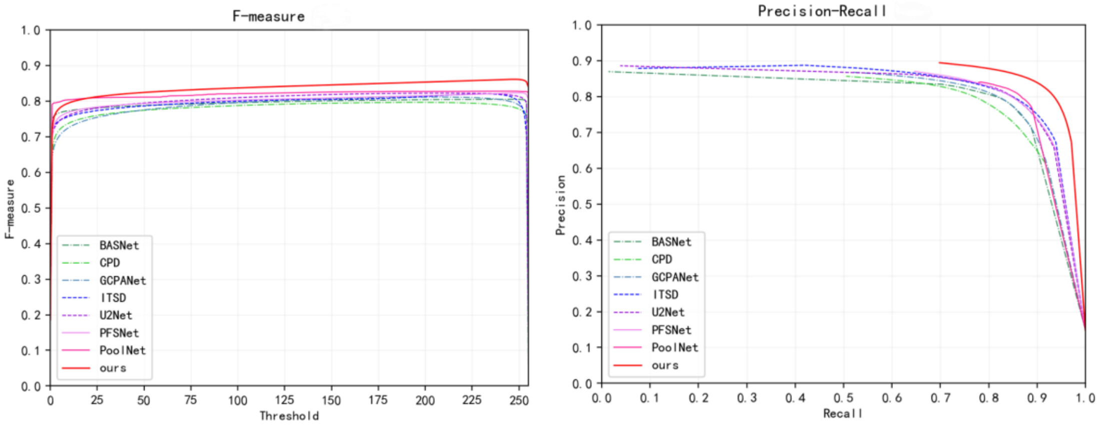

Fig.5 Curve chart

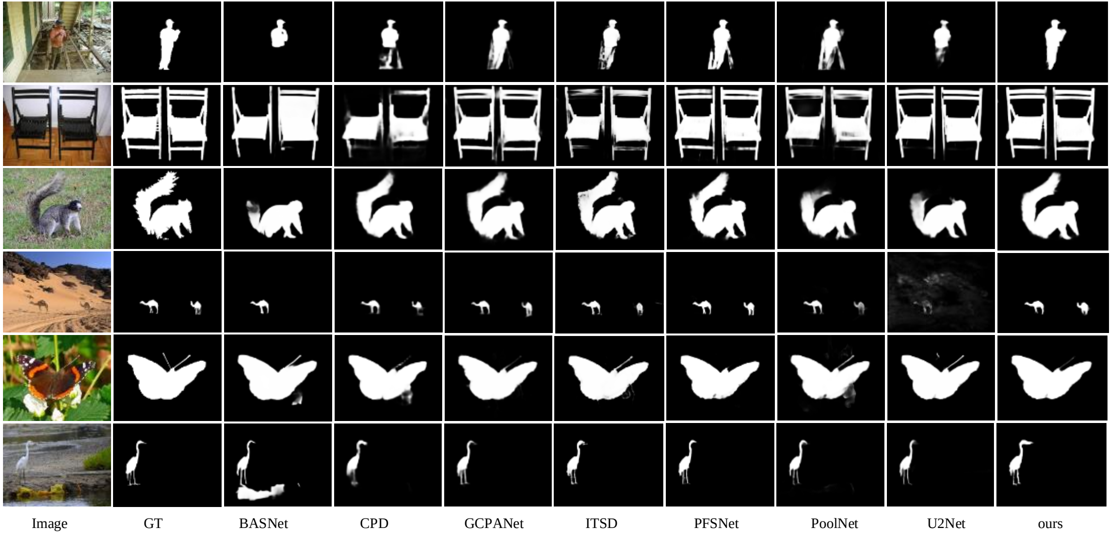

Fig. 6 Qualitative comparison

3.3.2 Qualitative Analysis

In order to compare the effectiveness of the algorithms more intuitively, Fig.6 provides a comparison of the visualization effects under different scenarios. It can be seen that compared with other algorithms, the proposed algorithm in this paper has a better ability to capture the semantic information and boundary details of the salient targets, and the completeness and boundary clarity of the prediction results are significantly improved. For example, in the first and fifth rows, the image scene is complex and occluded, and although other algorithms can locate the general area of the salient target, there is a lot of background noise in the prediction results, while the boundary of the proposed algorithm is relatively clearer and has less background noise. In addition, the second and fourth rows belong to the case of multiple disconnected salient objects, and the fourth and sixth rows belong to the case of small contrast between foreground and background colors and deviation from the center of the image, which are often difficult to locate the salient target and to clearly delineate the front background boundary area. From Fig.6, it can be seen that the proposed algorithm captures the fine boundaries of salient targets more accurately than other algorithms, and the prediction results are relatively closer to the true values.

3.4 Ablation experiments

To verify the effectiveness of each module, all ablation experiments in this section are conducted on the DUT-OMRON dataset. Tab.2 lists the results of the ablation experiments using the U-Net[27] model as the benchmark and gradually adding the feature fusion module FFM, the feature diversity enhancement module FDE, and the global information guidance module GIG. The visualization results of gradually adding each module in the text to the benchmark model are shown in Fig.7. It can be seen that after adding the feature fusion module to the benchmark network, using the same The F-measure value is increased from 0.742 to 0.770 and the MAE value is reduced from 0.089 to 0.076, which fully illustrates that the feature fusion module can improve the detection performance of the model and further improve the integrity of the feature extraction; subsequently, by adding the cascaded feature diversity enhancement module, the F-measure value increased from 0.770 to 0.798, and the MAE value decreased from 0.076 to 0.071, reflecting that the feature diversity enhancement module has the role of enriching spatial features, which makes the feature extraction ability of the model enhanced, and further iteratively refining significant features and suppressing background noise through the cascaded feature fusion module and feature diversity enhancement module; after adding the global After adding the global information guidance module, the tested F-measure value is improved from 0.798 to 0.820, and the MAE value is reduced from 0.071 to 0.054, indicating that the model can capture salient features more accurately, and the division between foreground and background is clearer, which alleviates the phenomenon of feature dilution.

Tab.2 Ablation experiments of different modules

Configurations				F-measure↑	MAE↓
Baseline	FDE	GIG	FFM		
√				0.742	0.089
√			√	0.770	0.076
√	√		√	0.798	0.071
√	√	√	√	0.820	0.054

Tab.3 Ablation experiment with different loss functions

Configurations	F-measure↑	MAE↓
FDE+GIG+FFM+L_{bce}	0.806	0.085
FDE+GIG+FFM+L_{hybrid}	0.820	0.054

In addition, in order to verify the effectiveness of the hybrid loss function, the saliency target detection model constructed in this paper was trained with the binary cross-entropy loss function and the hybrid loss function proposed in the paper, respectively, and the same training strategy was used during the training process, with all parameters kept consistent. According to the results in Tab.3, compared with the commonly used binary cross-entropy loss function, the F-measure value of the model tested with the hybrid loss function improved from 0.806 to 0.820 and the MAE value decreased from 0.085 to 0.054, reflecting that the hybrid loss function proposed in the paper has the effect of optimizing the significance target boundary and refining the significance prediction map. It further demonstrates the necessity of the hybrid loss function to improve the effectiveness of the algorithm in this paper. Fig.8 shows the visualization of the supervised graph during the model training.

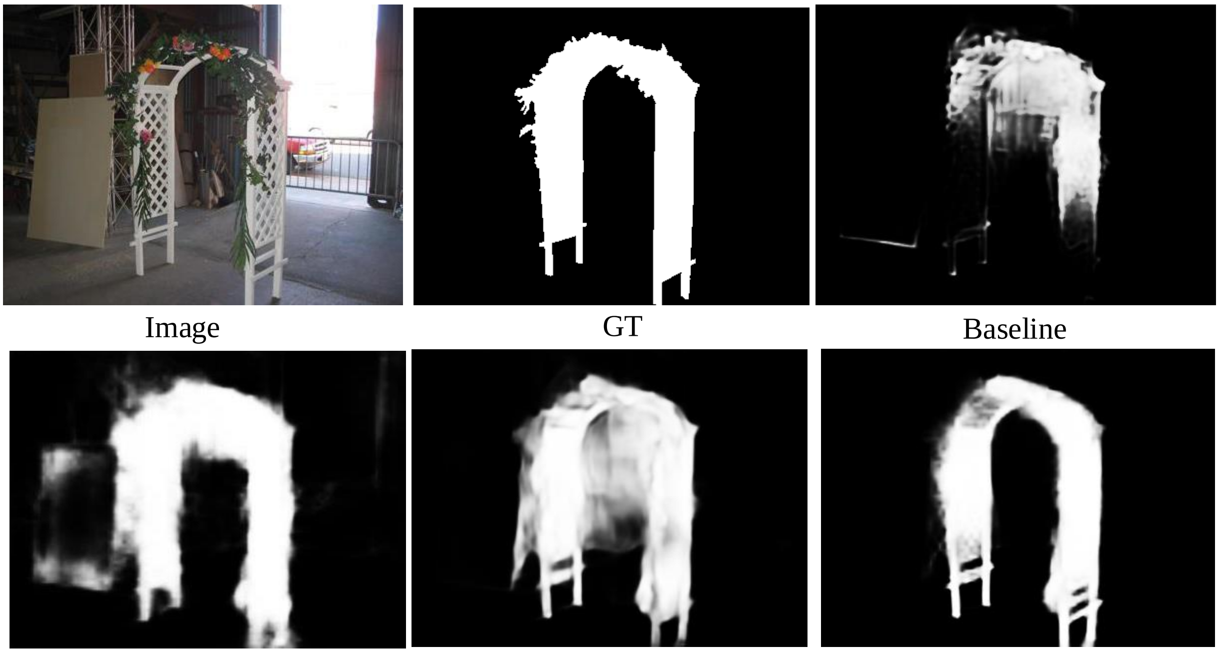
Image GT Baseline

Baseline + FFM Basleine + FDE Basleine + GIG

Fig.7 Visual comparison of each module

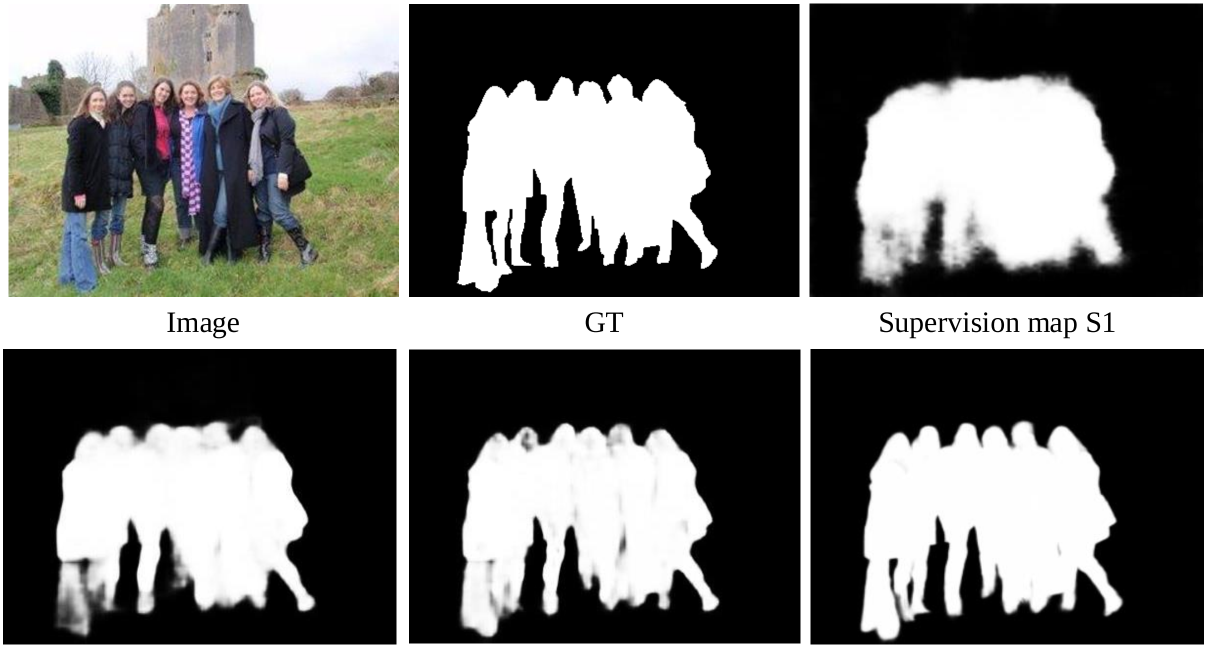
Image GT Supervision map S1

Supervision map S2 Supervision map S3 Supervision map S4

Fig.8 Visual comparison of deep supervision strategies

4. Conclusion

In order to improve the problem of insufficient completeness of the saliency prediction map, this paper firstly designs a feature diversity enhancement module formed by three different convolutional splices, which further captures many different types of spatial features based on the coarse features extracted by the backbone network; secondly, the constructed global information guidance module contains a coordinate attention mechanism and a global guidance flow, which is connected to the highest layer of the backbone network. The global guidance flow passes the high-level semantic information obtained by the coordinate attention module to each feature layer to improve the feature dilution problem generated by the top-down transfer process of the model; finally, the feature fusion module is used to aggregate the high-level features from the previous layer, the coarse features from the backbone network in the current layer and the global guidance flow from the global information guidance module to achieve the goal of suppressing background noise while using contextual information to accurately detect salient objects. In order to make the network pay more attention to the boundary information, this paper combines three loss functions with different applications to form a hybrid loss function, and uses a deep-supervised strategy to further sense the boundaries of salient objects. The effectiveness of the algorithm proposed in the paper and the necessity of each module are demonstrated by comparison experiments and ablation experiments. The comparison experiments with seven advanced algorithms on five publicly available datasets show that both F-measure metrics and MAE metrics achieve optimal or suboptimal results. In future work, we will consider combining edge detection branches to jointly train the network model to better improve the network performance.

References

[1] Cheng M M, Zhang F L, Mitra N J, et al. Repfinder: finding approximately repeated scene elements for image editing[J]. ACM transactions on graphics (TOG), 2010, 29(4): 1-8.
[2] Wang Q, Tang S, Zhai D, et al. Salience based object tracking in complex scenes[J]. Neurocomputing, 2018, 314: 132-142.
[3] Cheng M M, Mitra N J, Huang X, et al. Global contrast based salient region detection[J]. IEEE transactions on pattern analysis and machine intelligence, 2014, 37(3): 569-582.
[4] Liu R, Cao J, Lin Z, et al. Adaptive partial differential equation learning for visual saliency detection[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 3866-3873.
[5] Ma M, Xia C, Li J. Pyramidal feature shrinking for salient object detection[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2021, 35(3): 2311-2318.
[6] Qin X, Zhang Z, Huang C, et al. Basnet: Boundary-aware salient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 7479-7489.
[7] Wu Z, Su L, Huang Q. Cascaded partial decoder for fast and accurate salient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3907-3916.
[8] Wang Zhengwen, Song Huihui, Fan Jiaqing, et al.Salient Target Detection Network Based on Semantic Guided Feature Aggregation[J]. Acta automatica sinica,2021,48:1001-1010.
[9] Chen Z, Xu Q, Cong R, et al. Global context-aware progressive aggregation network for salient object detection[C]//Proceedings of the AAAI conference on artificial intelligence. 2020, 34(07): 10599-10606.
[10] Zhou H, Xie X, Lai J H, et al. Interactive two-stream decoder for accurate and fast saliency detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2020: 9141-9150.
[11] Darren Liu, Guo Jichang, Wang Yudong, et al.Multi-scale saliency target detection network based on attention mechanism[J]. Journal of xidian university (Natural Science Edition) ,2022,49(4):118-126.
[12] Liu J J, Hou Q, Cheng M M, et al. A simple pooling-based design for real-time salient object detection[C]//Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. 2019: 3917-3926.
[13] Qin X, Zhang Z, Huang C, et al. U2-Net: Going deeper with nested U-structure for salient object detection[J]. Pattern recognition, 2020, 106: 107404.
[14] Zhuge M, Fan D P, Liu N, et al. Salient object detection via integrity learning[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022.
[15] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.
[16] X. Ding, Y. Guo, G. Ding, and J. Han, "ACNet: Strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks," in IEEE Int. Conf. Comput. Vis., 2019, pp. 1911–1920.
[17] Wang P, Chen P, Yuan Y, et al. Understanding convolution for semantic segmentation[C]//2018 IEEE winter conference on applications of computer vision (WACV). Ieee, 2018: 1451-1460.
[18] Hou Q, Zhou D, Feng J. Coordinate attention for efficient mobile network design. arXiv 2021[J]. arXiv preprint arXiv:2103.02907, 2021.
[19] De Boer P T, Kroese D P, Mannor S, et al. A tutorial on the cross-entropy method[J]. Annals of operations research, 2005, 134(1): 19-67.
[20] Wang Z, Simoncelli E P, Bovik A C. Multiscale structural similarity for image quality assessment[C]//The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003. Ieee, 2003, 2: 1398-1402.
[21] Rahman M A, Wang Y.Optimizing intersection overunion in deep neural networks for image segmentation[C]//International symposium on visual computing. Springer, Cham, 2016: 234-244.
[22] Kanopoulos N, Vasanthavada N, Baker R L. Design of an image edge detection filter using the Sobel

operator[J]. IEEE Journal of solid-state circuits, 1988, 23(2): 358-367.

[23] Li Y, Hou X, Koch C, et al. The secrets of salient object segmentation[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2014: 280-287.

[24] Shi J, Yan Q, Xu L, et al. Hierarchical image saliency detection on extended CSSD[J]. IEEE transactions on pattern analysis and machine intelligence, 2015, 38(4): 717-729.

[25] Movahedi V, Elder J H. Design and perceptual validation of performance measures for salient object segmentation[C]//2010 IEEE computer society conference on computer vision and pattern recognition-workshops. IEEE, 2010: 49-56.

[26] Yang C, Zhang L, Lu H, et al. Saliency detection via graph-based manifold ranking[C]//Proceedings of the IEEE conference on computer vision and pattern recognition. 2013: 3166-3173.

[27] Ronneberger O, Fischer P, Brox T, et al. U-net: Convolutional networks for biomedical image segmentation[C]//International Conference on Medical image computing and computer-assisted intervention. Springer, Cham, 2015: 234-241.