Pattern Recognition 171 (2026) 112118

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

IFA: Illumination-aware feature aggregation model for salient object detection

Miao Li [a,1], Hongyun Zhang [a,1,*], Kecan Cai [a], Witold Pedrycz [b], Duoqian Miao [a], Ying Gao [c]

[a] School of Computer Science and Technology, Tongji University, Shanghai, 201804, China
[b] Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB T6G 2G7, Canada
[c] School of Management and Engineering, Capital University of Economic and Business, Beijing, 100070, China

ARTICLE INFO

Keywords:
Salient object detection
Retinex theory
Illumination map
Feature extraction

ABSTRACT

Significant progress has been made in salient object detection. Most methods extract information directly on RGB features result in a loss of semantic information. When others introduce spatial information by using depth, difficult-to-access depth and the imbalance of depth have brought serious challenge. In this paper, we introduce illumination map based on Retinex theory inspired by low-light enhancement to SOD. Easy-to-obtain illumination can not only supplement semantic information, but also one can avoid the problem of unbalanced object depth. We design the dual-scale progressive fusion module, which can aggregate the illumination by proposed cross-scale information fusion. We embed it to design an illumination-aware model, where we also build a tailored multi-scale feature connection prediction network to obtain salient maps. Through experiments completed on five datasets, our method has demonstrated excellent detection ability and provides a novel solution for SOD,especially in underwater and low-light datasets our method still achieves outstanding results.

1. Introduction

Salient object detection (SOD) has become an important research field in artificial intelligence. It is essential to the development of computaer vision. Benefiting from the good performance of feature extraction of convolution neural networks, more and more SOD methods based on deep learning have been proposed, while showing superior performance. Very often, many researchers use a single RGB image without any auxiliary input information as the only given input to get a salience map [1]. These methods build an end-to-end deep neural network for SOD. Although, in general, the methods with single input can produce good results, it cannot obtain satisfactory results in many complex scenes, such as cluttered objects, similar objects to background and nonuniform lightness. In particular, a single RGB-based model is not good at forming the difference between the object and the background, these spatial information are particularly important in the task of salient object detection. Therefore, some methods [2] are designed to supervise the edge information of the salience object by learning the edge feature of the origin image. Recently, more methods that use the depth map of the RGB image to assist the salient object detection are constantly proposed, and achieved good results.

However, depth maps have two disadvantages for the task of SOD. First, depth maps [3] are not sufficient to provide full geometric and spatial information because of disequilibrium of object depth. As shown on the right hand side of Fig. 1, the car head and rear have different depths of field, so the depth image shows different characteristics. But the salient object detection task pays more attention to the ***whole*** object itself rather than the difference of the depth information. Second, it is difficult to obtain the effective and accurate depth maps. Because the acquisition of depth maps often requires a large amount of equipment and complex computation or is easily affected by the lighting and noise of the scene. Despite the advent of portable depth-sensing cameras, there is still room for improvement in the quality of the obtained depth maps.

To solve these kinds of problems, illumination map, as one of the most easily obtained information in images, is gradually used in the task of RGB-T SOD in recent years [4]. In the existing deep learning methods, low-light scenes are often encountered in different tasks. At present, for low-light scenes or unbalanced illumination scenes, extracting illumination maps is the key to solve many downstream tasks. The main reason lies in the fact that high quality illumination maps usu-

* Hongyun Zhang is the corresponding author and joint first author.
E-mail address: zhanghongyun@tongji.edu.cn (H. Zhang).
[1] Joint first authors.

https://doi.org/10.1016/j.patcog.2025.112118
Received 26 June 2024; Received in revised form 25 June 2025; Accepted 2 July 2025
Available online 15 July 2025

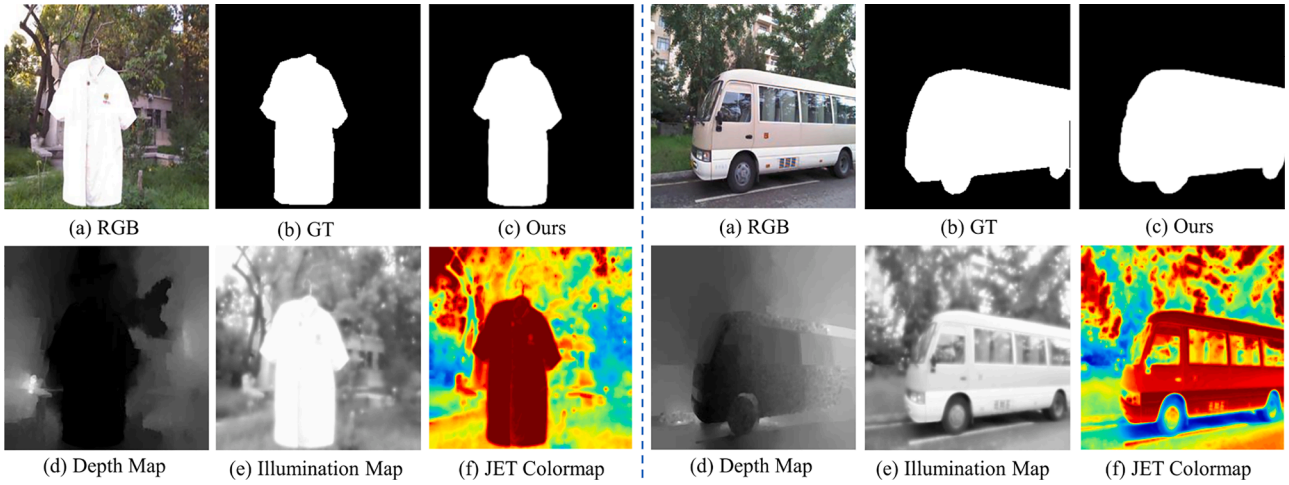


Fig. 1. Examples of images and auxiliary maps. Image (a), (b), and (d) respectively represent the RGB image, its corresponding Groundturth image, and depth map. (c) Represents the result obtained from our model, while (e) illustrates the illumination map of the image. Finally, (f) represents the representation of the illumination map in the JET color space.

ally reflect the intensity and ability of the reflected illumination of the object. Cong et al. [5] convert illumination feature to one global parameter to guide the fusion of RGB and infrared features. These methods all make use of the average parameters of the global illumination, and this parameter cannot represent local illumination. Inspired by these methods, we use the illumination maps as an auxiliary input for salient object detection.

Inspired by these study, we propose a novel illumination-aware model (IFA) by introducing the illumination map based on ***Retinex theory*** [6]. IFA uses easily accessible illumination maps as auxiliary inputs to obtain information about how the image reflects light, rather than just representing the intensity of the light. We use feature maps instead of global illumination parameters to assist the SOD task As shown in Fig. 1(f), the illumination map contains light information for different objects, which smooths out complex backgrounds and reduces information redundancy. However, it is difficult for one model to extract both the illumination map based on Retinex theory and the salient area, we introduce an independent decomposition model to obtain the illumination map. Another advantage rather than fixed parameters is that the fusion of illumination maps and RGB images can enhance the perception of light, thus reducing the impact of light on the model, such as low-light scenes.

Another important consideration is the fusion of illumination map and RGB image. In Fig. 1, different from depth map, the illumination map is less affected by the complex background. In particular, in Fig. 1(f), the illumination maps can significantly highlight the spatial position information of the object. However, due to many methods use the same backbone to extract features currently, the features highlighted by the illumination maps and RGB images do not align well. Therefore, we design a cross-scale feature fusion structure to fuse illumination information of different scales.

In particular, firstly, we use the DecomNet based on Retinex theory to form the illumination maps of the image, which is an excellent low-light image enhancement algorithm with a series of CNN layers. Secondly, we adopt the ResNet as the backbone to extract the RGB and illumination features with different scales. To better solve the problem of feature alignment when fusing illumination maps and RGB features, we propose a dual-scale progressive feature fusion module (*DSPM*), and use it to fuse the low-scale RGB feature and high-scale illumination features by proposed cross-fusion convolution module. Since high-scale illumination feature contain more location information, which can effectively matches the semantics of low-scale RGB features. Thirdly, to make better use of the spatial position information, three fused features on three scales by DSPM were fed into the proposed tailored multi-scale feature connection prediction network (*TMSN*), which is used to get spatial position information by different scale features. TMSN mainly is an U-Net structure and uses the pooling layer to concatenate these fused features at the lowest scale, and uses skip-connection to obtain contextual features. In this way, we achieve the task of salient object detection by fusing illumination maps.

In summary, this study offers four major contributions:

- We apply accessible illumination maps based on Retinex theory for the task of salient object detection. The main aim is to extract the geometric and spatial information of the objects better by fusing illumination features.
- We propose a dual-scale progressive feature fusion module (DSPM), which consists of Dual-scale Feature Extension Block and Dual-branch Feature Aggregation Block to fuse illumination maps and RGB features on different scales, and this module can be used to fuses high-level and low-level features of images.
- We build a tailored multi-scale feature connection prediction network (TMSN) that can extract multi-level and multi-scale features. Combined TMSN and DSPM, an illumination-aware feature aggregation network (IFA) is built, which can be effectively applied to SOD tasks to datasets of different quality.
- We introduce a feature reduction module (FR) to further improve model performance and efficiency. Compared to the proposed model without feature reduction module, the proposed method achieves 78% reduction in the number of parameters, while simultaneously attaining average 9.12% MAE performance improvement on five datasets.

In summary, an original approach for salient object detection task is proposed, which uses illumination maps based on Retinex theory as auxiliary input to detect salient objects. This innovation was inspired by low-light image enhancement tasks. This can effectively avoid the impact of complex, low-quality scenes. In order to better extract and fuse illumination features, we design a multi-stage and multi-scale illumination-aware network (IFA) that can combine context to effectively segment the edge of salient objects.

The organization of this study is as follows. Firstly, the related work is presented in Section 2. Secondly, the proposed models IFA is extensively described in Section 3. The experimental results are showcased in Section 4. The expended experiments are presented in Section 5, which includes the results on underwater and low-light image datasets. Lastly, the conclusions and future work are provided in Section 6.

2. Related works

2.1. Salient object detection

During the past decades, salient object detection has been one of the research hotpots exhibiting significant development potential. Early salient detection algorithms focused on studying the internal structure of images, aiming to detect images through texture and edge information.

With the development of deep learning, methods using convolutional neural networks are increasingly being proposed. Liu et al. [7] analyzed the role of context information in SOD and designed a pixel-based context attention network. Deng et al. [8] designed a recurrent residual refinement network to use the multi-level features, and at the same time the residual difference between the result and ground truth is used to learn the salient objects.

In recent years, more and more studies pay attention to more advanced models and ideas with the improvement of method performance. Liu et al. [9] focused on the role of pooling layer in SOD tasks, and proposed a pooling-based feature aggregation network (PoolNet), which was gradually improved to use edge information for training (PoolNet+). Instead of using pixel-based detection, Chen et al. [10] designed a model that could learn features at different levels, effectively integrating low-level appearance features and high-level semantic features. Zhao et al. [11] introduced the idea of gated network, which effectively solved the disparity brought by traditional encoder and decoder network. Wu et al. [12] designed a generation model to directly generate training data, which greatly reduced the cost of tasks. Zhou et al. [13] designed an interactive dual-stream encoder to learn the correlation between saliency detection maps and corresponding contour maps, enabling the reverse learning of model parameters. This method effectively detects object edges while achieving outstanding performance. Li et al. [14] focuses on both salient area and non-salient area detection, and proposes a novel complementary perceptual attention network. Zheng et al. [1] constructed a learning network based on just distillation, which can effectively enhance the generalization ability of student networks and achieve saliency goal detection tasks. Wang et al. [15] also used a weakly supervised network to achieve salient object detection, and they used graffiti as guidance information. Yi et al. [16] designed a progressive fusion framework, which is a gated fusion model that aims to reduce the redundancy of image features

However, it is worth noting that most methods in the field still rely solely on a single RGB image as the input for saliency detection. While they have shown promising results, they often struggle to effectively highlight salient objects in more complex scenes, which remains a common limitation among existing approaches.

2.2. Auxiliary maps for SOD

In recent years, salient object detection based on RGB images has shown excellent performance. However, challenges still exist in similar background environments, low-quality conditions, and complex scenes. Therefore, exploring the role of various auxiliary features has become a key factor in improving the performance of salient object detection models.

Depth maps, as they contain the 3D information of the image, can be effectively applied to salient object detection tasks. Fan et al. [17] skillfully combined cross-modal techniques and low-quality depth feature filtering to significantly enhance the role of depth information in saliency object detection, all the while efficiently utilizing the rich 3D information from depth maps. Chen et al. [18] proposed a weakly supervised network based on graffiti annotations and designed a two-model framework to learn complementary information. On the other hand, Liu et al. [19] devised a remarkable triple transformation embedding module, incorporating a weight-shared encoder. This ingenious module successfully integrated depth information into RGB features, achieving an impressive effect of multimodal fusion. Zhu et al. [4] designed a semi-supervised learning network based on bimodality, which utilizes dynamic adjustment and active expansion strategies to explore the model's learning capabilities In addition to depth maps, edge estimation methods have also been harnessed in salient object detection. Yang et al. [2] employed a edge-preserved connectivity-based approach to obtain initial saliency detection maps, and then ingeniously fused them with edge information. By leveraging edge detection techniques, they succeeded in significantly enhancing the overall performance of salient detection.

While the methods mentioned above can indeed supplement RGB images with additional information, the challenge lies in obtaining precise depth maps, particularly in low-quality and complex scenes, which becomes a bottleneck for RGB-D tasks. Moreover, edge estimation methods have limitations in providing sufficient object information. To tackle this issue, some novel approach based on RGB-T has been proposed, introducing thermal (T) information for salient object detection [5]. However, acquiring thermal maps proves to be more difficult that necessitates expensive equipment. Hence, the search for a suitable auxiliary input becomes paramount in enhancing overall performance in these scenarios.

2.3. Model efficiency and tasks survey

At present, a large number of salient object detection methods have been proposed, and there have been related reviews. Zhou et al. [20] introduced a survey of depth-based methods, describing the relevant datasets and the theory of salient object detection with RGB-D in detail. These approaches explore the main motivation of previous work, including what data is able to assist task completion? What can be done to make the model more efficient?

This paper discusses how to use auxiliary input to help salient object detection task, and at the same time, we try to meet the efficiency requirements of the model as much as possible. In the previous methods, the study of model structure plays an important role in the application of deep learning tasks. Jing et al. [21] reprogram a pre-trained GNN, without amending raw node features nor model parameters, to handle a bunch of cross-level downstream tasks in various domains. Yang et al. [22] decompose the model into independent structures and then reassemble the custom model according to the function of the module, aiming to achieve the knowledge transfer task. Yang et al. [23] use modularization and assemblability of knowledge to explore *Knowledge Factorization*, and use modular structure to explore the influence between different modules. These methods are discussed from the model structure level to explore the effectiveness and functionality of model components.

3. Methodology

3.1. Architecture overview

In this section, we first introduce the overall structure of the proposed method. Our framework is divided into three parts: feature generation part, information aggregation part and salient object prediction part. As is shown in Fig. 2, we use a structure to obtain the illumination maps based on Retinex theory. The module is called Decomposition Network (*Decom-Net*), whose output is a map with single channel.

Subsequently, the feature extraction, such as VGG and Res-Net, are used as the backbone, which processes the original image and illumination map to obtain the intermediate features with different resolutions, which are denoted as $\{X_i\}_{i=1}^{4}$ and $\{I_i\}_{i=1}^{3}$.The two backbones use the same structure, and the number of output features is different.

Specifically, once the RGB image has been extracted by backbone network, the channel number of four resulting feature ($X_1 - X_4$) are: 256, 256, 512, 1024. To reduce the feature redundancy and alleviate complexity of the model, we introduce a feature reduction module. We use the convolution with 3x3 kernels to change the numbers of feature

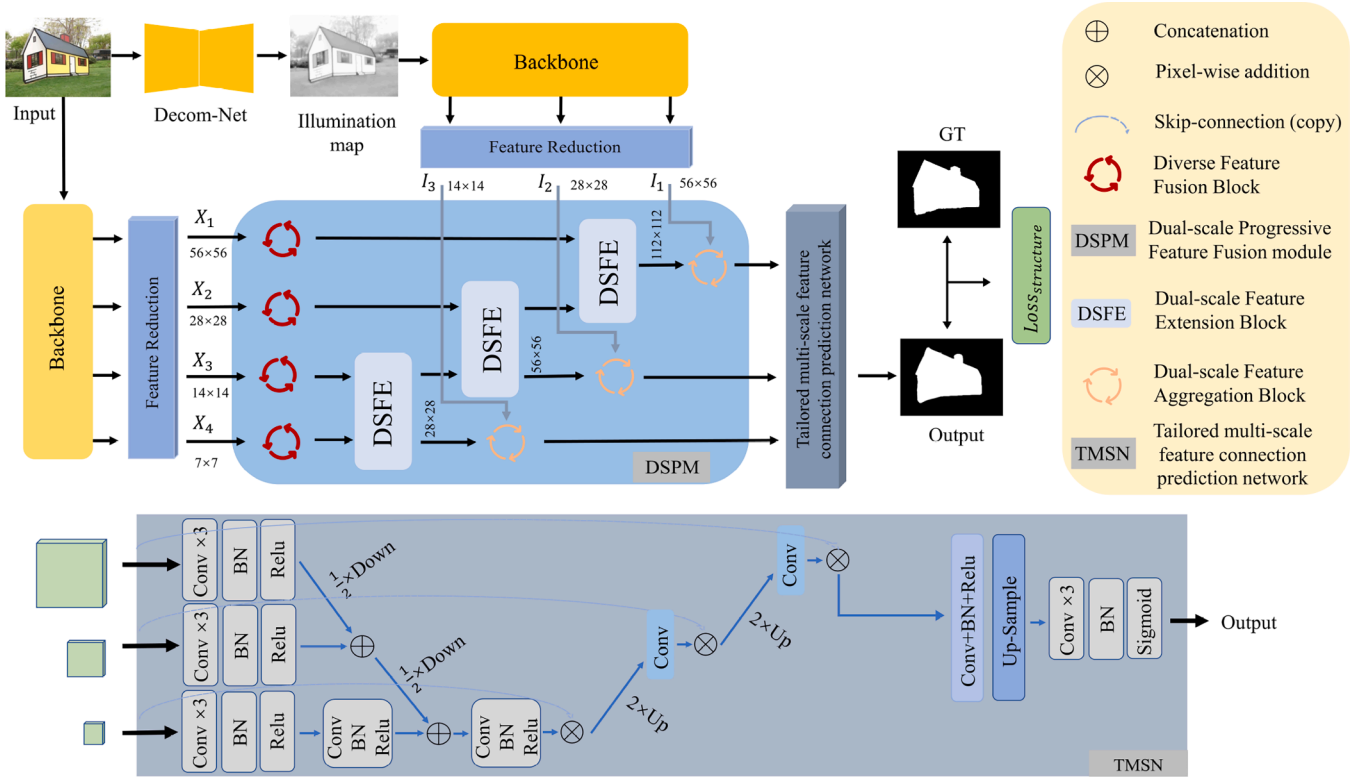


Fig. 2. The overall structure of the proposed model.

channel (the channels are reduced to 16, 32, 64, 128), and specific experimental results are introduced in detail in Section 4.

Finally, we fuse the features of the illumination map and the features of the input image, and get the prediction results. Fig. 2 shows the proposed model named illumination-aware feature aggregation network(IFA).

3.2. Illumination maps extraction

Retinex theory states that the illumination map depends on the inherent properties of different objects, and contains intensity information of the light at different positions and the reflection ability of the different objects in the image to light. Among them, the extraction and fusion of light intensity into RGB can strengthen the light intensity and make the light information of the low-illumination image more abundant.

According to Retinex theory, an image can be represented by the following expression: $L(x, y) = R(x, y) * I(x, y)$. Where L represents the original image, R and I represent the reflection and illumination map of L, and $*$ represents multiplication. Retinex theory has achieved good performance in low-light image enhancement tasks [24]. Inspired by RetinexNet [6], we build a structure named *Decom-Net* to obtain illumination maps, the specific structure is shown in Fig. 3. The illumination maps are trained from the low-light image dataset [6] by transfer learning, the reason is that, compared to low-light image enhancement tasks, salient object detection models cannot effectively extract illumination maps based on Retinex theory.

More specifically, *Decom-Net* is a separately trained network, which has an input and two outputs. Since the low-light image enhancement based on Retinex can better process the illumination information, we use LOL-dataset [24] (a generic paired low-light image dataset) to learn the representation of Decom-Net. As shown in Fig. 3, the features of the first three channels defined as $R = Concat\left(\sum_{i=1}^{3}\{Output_i\}\right)$ constitute the reflection map of the image in the output results of the *Decom-Net*, and the feature of fourth channel is regarded as the illumination map

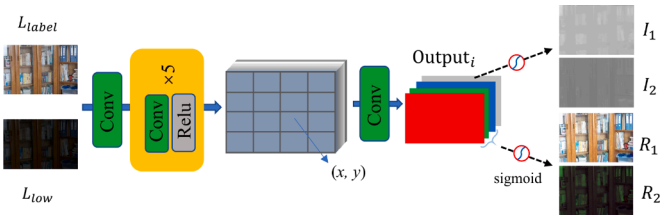


Fig. 3. Decomposition module based on Retinex theory. Decom-Net generates a four-channel feature, with the first three channels representing the reflection map and the fourth channel representing the illumination map. During training process, both low-light images and normal-light images are fed into the model, and the two obtained reflection maps are used to calculate the loss. The final result is the illumination map.

$I = Output_4$. According to Retinex theory, the loss function of our Decom-Net can be designed as follows:

$$Loss = \sum_{j=1,2}\left\|R_j * I_j - L_j\right\|_1 + \beta\sum_{j=1,2}\left\|\nabla I_j * e^{-10\nabla R_j}\right\|_1 + \alpha\sum_{j=1,2}\left\|R_j * I_{2-j} - L_{2-j}\right\|_1, \quad (1)$$

where L_1,R_1, I_1 and L_2,R_2,I_2 represent the normal/low-light image, reflection map and illumination map, and $\|\cdot\|_1$ represents the L1 norm loss function. Based on Retinex, the values of α and β are typically set to 0.1 and 0.05. ∇ defined as the average value of the gradient in the portrait or landscape orientation of the image. The logic behind the equation is that we get the same illumination from low/normal-light image in theory.The specific realization is to calculate the difference (such as gradient) between the illumination maps (I_1 and I_2), and the Eq. (1) is used to verify whether the multiplication of the maps after decomposition ($R_j * I_j$) conforms to Retinex theory.

Algorithm 1 Feature reduction module.

Input: RGB features x
Output: Reduced features
Normalization x, $n \leftarrow 4$
Compute the features of different backbone layer x_i of x, $i \in \{1, 2, 3, 4\}$
for $k = 1, m = 1$; $k \leq n$; $k++$ **do**
  while $m \leq 2$ **do**
    Convolution layer with 3×3 kernels, The output channel is reduced by twice
    Normalization x_i, Relu x_i, m++
  end while
end for
if x is RGB feature **then**
  return f_1, f_2, f_3, f_4
else
  return f_1, f_2, f_3
end if

3.3. Feature reduction

Our FR module is mainly used to reduce the number of feature channels. According to our experiments, we found that excessive deep feature channels could not improve the performance of SOD, but increased lots of model parameters. Therefore, we reduced the number of feature channels by using a two-layer convolutional neural network to facilitate the reduction of redundant information. The specific algorithm of the module is shown in Algorithm 1.

3.4. Dual-scale progressive feature fusion module

As aforementioned, this part mainly includes three feature processing blocks: Diverse Feature Fusion Block (*DFF*), Dual-scale Feature Extension Block (*DSFE*) and Dual-scale Feature Aggregation Block (*DSFA*).

3.4.1. Diverse Feature Fusion Block

For the features $\{X_i\}_{i=1}^4$ obtained by the different depth neural network layers of backbone, we adopt a three-branch structure to generate a more discriminative feature representation. Let X denote any member of the feature set $\{X_i\}_{i=1}^4$, in the three branches shown in Fig. 4(a), we respectively use the single convolution layer, the convolution layer with the parameter: $dilation = 1$, and a convolution with batch normalization and Relu function, which can be expressed as:

$$Y = conv(conv\{X\} \oplus conv\{X, dilation = 1\} \oplus Relu\{BN(Conv\{X\})\}). \quad (2)$$

where Y represent the output of the *DFF*, BN denote as the Batch Normalization layer. The input and output of *DFF* have the same size and channel because of the last layer of convolution.

3.4.2. Dual-scale Feature Extension Block

Dual-scale Feature Extension Block is mainly designed to make full use of the feature information at different scales.It is a dual-input and dual-output block, and its output is twice the size of the input. *DSFE* is mainly divided into two stages: feature extension stage and feature aggregation stage.

DSFE is a form used for block progressive stacking, where the feature with small size in the output of the former is one of the inputs to the next block. Specificity, let $\{Y^i\}_{i=1}^4$ denote the features from four *DFF* blocks: $DFF(X_1)$, $DFF(X_2)$, $DFF(X_3)$ and $DFF(X_4)$. For Y^n with H4×W size and Y^{n-1} with 2H×2W size, we first expanded their sizes from H×W and 2H×2W to H×W, 2H×2W and 4H×4W by cross-fusion of up/down-sample. The features obtained at the first stage can be expressed as:

$$f^{4size} = Up_2(Y^{n-1}),\ f^{2size} = conv(Y^{n-1})Up_2(Y^n),$$
$$f^{size} = Dw_2(Y^{n-1})conv(Y^n). \quad (3)$$

Where $Up_j(\cdot)$ and $Dw_j(\cdot)$ denote the up-sample with interpolation (*j) and Down-sample expressed as convolution with the parameter: $stride = j$. f^{size}, f^{2size} and f^{4size} represent the outputs of the first stage with different scales.

The second stage is the reverse process to the one completes at the first stage in terms of the number of input and output features, except that the feature sizes of the output are still doubled. We don't use the pooling layers throughout the *DSFE* block. Given the features obtained in the first stage: $\{f^{size}, f^{2size}, f^{4size}\}$, the final output is expressed as follows, where $\otimes$ denotes concatenation and $m \in \{1, 2, 3\}$.

$$\begin{cases} output' = conv^2\{conv^2(I^m) \otimes Dw^2(conv(F^m))\}, \\ output'' = conv^2\{conv^2(F^m) \otimes Up^2(conv(I^m))\}, \\ output = output' \otimes Dw^2(output''), \end{cases} \quad (4)$$

The *DSFE* block uses Y^4 and Y^3 as input for the first processing and on the second *DSFE*, we replaced the smaller size output of the last *DSFE*

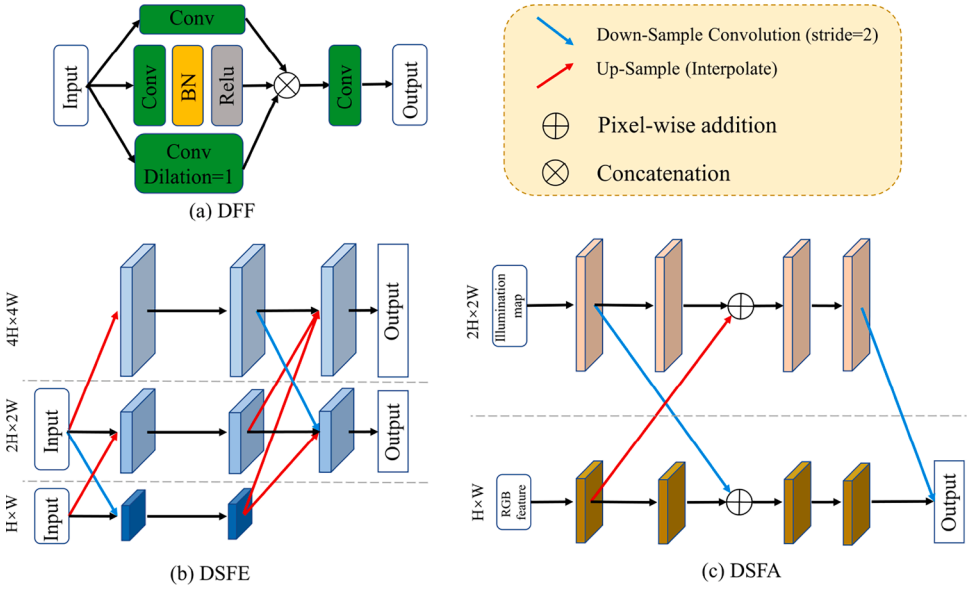


Fig. 4. The proposed feature extraction modules.

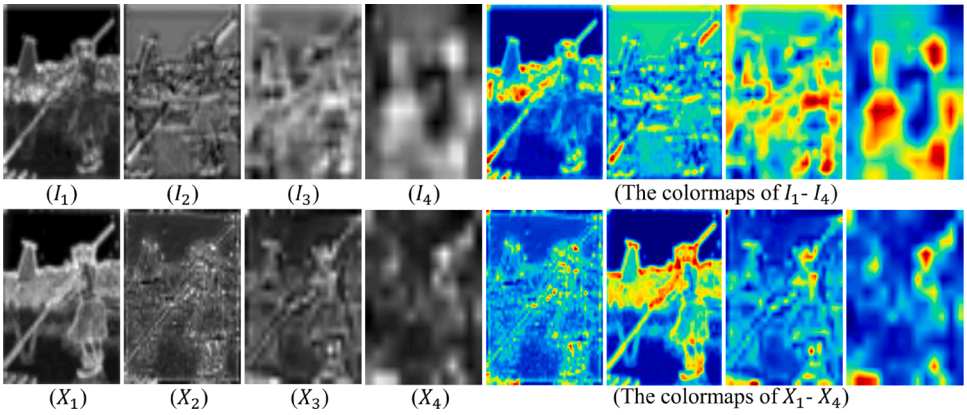


Fig. 5. Feature maps captured at different levels of backbone.

with Y^{n-1} while fusing Y^2. The larger size feature of each block is for subsequent fusion with illumination maps. Since the illumination maps in this paper is designed with three outputs from backbone, we only consider the larger feature out^l in the last *DSFE* block.

3.4.3. Dual-scale Feature Aggregation Block

Salient object can be transferred to feature maps of different scales. In this paper, considering that the illumination maps are the characteristic representations being different from RGB images, we propose a two-scale down-sample feature aggregation block (*DSFA*) to absorb illumination map.

Our method uses larger scale feature of image to fuse smaller scale illumination map and the specific structure is shown in Fig. 4. Fig. 5 shows the visualization of feature extraction of light component and RGB image by the same model. The results show that the smooth illumination map and the RGB image with complex information are not synchronized in feature extraction (The same depth indicates different details). Compared to RGB images, illumination features with a larger number of channels often contains a amount of feature redundancy. Feature reduction module effectively reduces the number of channels of illumination map features. In addition, we designed a *DSFA* module to integrate different scale features.

Let F^1, F^2, F^3 denote the features obtained by *DSFE* from large to small and I^1, I^2,I^3 represent the illumination maps, where $\{F^m\}_{m=1}^3$ is twice the size of corresponding $\{I^m\}_{m=1}^3$. Therefore, the output of *DSFA* can be expressed as:

$$\begin{cases} output' = conv^2\{conv^2(I^m)Dw^2(conv(F^m))\} \\ output'' = conv^2\{conv^2(F^m)Up^2(conv(I^m))\} \\ output = output' Dw^2\left(output''\right) \end{cases} \tag{5}$$

where $conv^{\times 2}(\cdot)$ represents two layers of convolution. In this way, we obtain the depth characteristics of the three scales (m belongs to the set $\{1,2,3\}$).

3.5. The structure of TMSN

The third part of the proposed model is tailored multi-scale feature connection prediction network (*TMSN*), which is also the last part of the algorithm. The overview of this part is shown in Fig. 2.

Given three inputs in descending order (for feature scale): X_1, X_2, X_3 these features are first preprocessed with a feature extraction layer (L'), which includes three 3×3 convolution, a batch normalization and a Relu function, where X_3 use two L'. Then we get the concatenation of the features (C) in the smaller scale, which contsains the larger receptive field. The processed feature can be expressed as the equation: $C = Down^2\left[Dw^2\left(L'(X_1)\right) \oplus L'(X_2)\right] \oplus L'(X_3)$, where $\oplus$ denotes the pixel-wise addition. Subsequently, we gradually restore the deep features to the required size and the output can be expressed as:

$$\begin{cases} out^k = X_k \otimes conv\left(Up^2\left(out^{k-1}\right)\right) \\ out^0 = L''(C) \end{cases}, \quad k \in \{1,2,3\} \tag{6}$$

where $L''(\cdot)$ represents the feature extraction with a 3×3 convolution, a batch normalization and a Relu function. Finally, we can obtain the salient map using a series of convolution flows, which is shown in following equation and the output and input of proposed model have the same size: $output = L'\left(Up^2\left(out^3\right)\right)$.

3.6. Loss function

In this study, we use the weighted binary cross entropy (BCE) loss and weighted intersection over union (IoU) loss. Since the single BCE loss ignores the overall structure of the image, we use the weighted BCE loss as a supplement and the overall loss function is shown as follows. $wBCE$ and $wIoU$ are described in detail in Ref. [13].

$$Loss = Loss^{wBCE} + Loss^{wIoU}. \tag{7}$$

4. Experiments and analysis

4.1. Experimental settings

Implementation details. The proposed model is implemented based on Python 3.7.1 with Pytorch framework. The experiments are performed on Ubuntu system with Nvidia RTX3090 GPU (24G RAM). The Adam optimizer with the parameters $\beta_1 = 0.9$, $\beta_2 = 0.1$ is used to train the experiments. We trained the model for a total of 120 epochs and adopt the decaying learning rate which express as $lr = init_{lr} * 0.1^{(epoch//300)}$, where $init_{lr} = 10^{-4}$, $epoch$ denotes the current epoch, we also update the learning rate every 300 epochs. Our model is trained based on the DUTS-TR dataset ($batchsize = 24$). Firstly, we utilize the decomposition module to obtain the illumination component of the image, and then generate a single saliency prediction map.

Datasets. We use a variety of datasets to validate proposed method, including the datasets listed in Ref. [25] (ECSSD, PASCAL-S, DUT-OMRON, HKU-IS, DUTS-TE), which contained 1000, 850, 5168, 4447 and 5019 test images separately.

Evaluation criteria. We use widely-used metrics such as mean F-measure score (Fb, larger is better), mean absolute error (MAE, smaller is better), and mean E-measure score (mE, larger is better) to verify the validity of the model and compare it with previous state-of-the-art methods. The mean F-measure score can be expressed by the following equation:

$$F_\beta^{mean} = \frac{1}{N}\sum_{i=1}^{N} \frac{\left(1+\beta^2\right) Precision \times Recall}{\beta^2 Precision + Recall}, \tag{8}$$

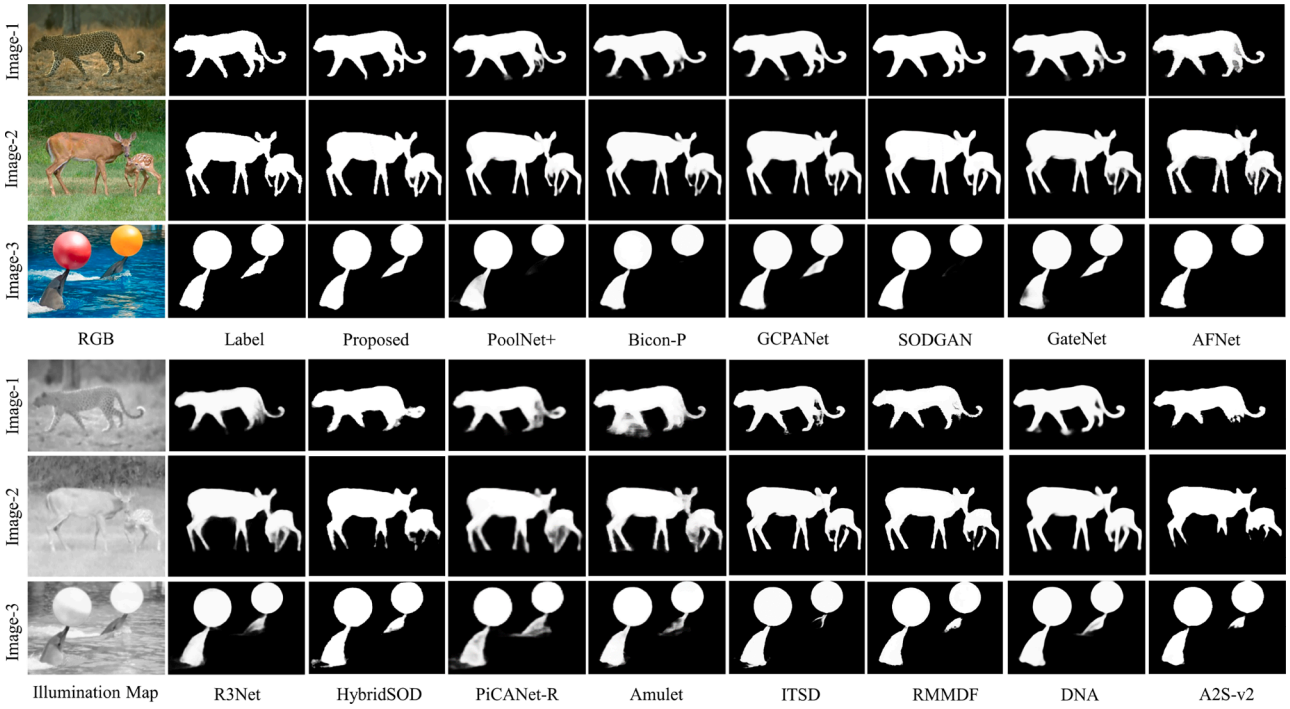


Fig. 6. Subjective comparison of the experimental results among different methods.

Table 1
Performance comparison between proposed method and the other methods on 5 datasets. The bold is the best result, *blue* is the next best result and the green is the third level.

	ECSSD			DUTS-TE			PASCAL-S			HKU-IS			DUT-OMRON		
$Methods_{year}$	MAE	Fb	mE	MAE	Fb	mE	MAE	Fb	mE	MAE	Fb	mE	MAE	Fb	mE
$Amulet_{17}$ [27]	.0592	.8727	.9104	.0846	.7015	.8156	.0980	.7672	.8349	.0521	.8453	.9071	.0977	.6677	.7934
$PiCANet\text{-}R_{18}$ [7]	.0465	.8902	.9252	.0498	.7887	.8747	.0768	.7922	.8536	.0433	.8658	.9161	.0648	.7233	.8294
$R3Net_{18}$ [8]	.0563	.8905	.9132	.0661	.7560	.8452	.1050	.7648	.8069	.0483	.8660	.9092	.0711	.7188	.8263
$AFNet_{19}$ [31]	.0418	.9059	.9355	.0453	.8107	.8909	.0760	.8073	.8662	.0358	.8878	.9343	.0573	.7397	.8424
$GCPANet_{20}$ [10]	***.0356***	.9124	.9420	*.0378*	***.8358***	.9067	***.0678***	***.8148***	.8763	.0316	.8976	.9419	.0566	.7488	.8468
$GateNet_{20}$ [11]	.0418	.9034	.9310	.0448	.8139	.8893	.0741	.8104	.8643	.0360	.8891	.9322	.0613	.7292	.8357
$ITSD_{20}$ [13]	*.0346*	*.9208*	***.9470***	.0408	*.8390*	*.9125*	.0728	*.8194*	***.8774***	***.0307***	***.9035***	.9472	.0608	***.7672***	***.8640***
$PoolNet+_{22}$ [9]	.0388	.9141	.9400	***.0397***	.8297	***.9011***	.0795	.8107	.8617	.0321	.8990	.9422	.0554	.7491	.8476
$Bicon\text{-}P_{22}$ [2]	.0358	.9187	*.9490*	.0418	.8262	.8888	.0772	.8130	.8624	.0351	.8956	.9282	.0582	.7413	.8359
$SODGAN_{22}$ [12]	.0389	.9004	.9438	.0530	.7808	.8870	***.0700***	.8071	***.8823***	.0324	.8909	***.9492***	.0768	.7121	.8333
$CANet_{22}$ [14]	.0492	.8844	.9257	.0556	.7659	.8703	.0862	.7894	.8612	.0401	.8707	.9296	.0705	.6995	.8298
DNA_{22} [28]	.0427	.8941	.9345	.0464	.7869	.8881	.0836	.7938	.8576	.0360	.8735	.9362	.0630	.7204	.8441
$RMMDF_{22}$ [25]	.0422	***.9193***	.9355	.0494	.8198	.8904	.0832	.8133	.8587	.0334	*.9118*	.9424	*.0532*	*.7766*	*.8700*
$A2S\text{-}V2_{23}$ [26]	.0441	.9144	.9366	.0468	.8143	.9010	.0796	.8125	.8712	.0365	.9014	*.9494*	.0609	.7500	.8636
$AccoNet_{23}$ [29]	.0415	.9076	.9365	.0557	.7925	.8808	.0805	.8006	.8592	.0405	.8841	.9297	.0604	.7606	***.8640***
$FPSI_{24}$ [30]	.0360	.9070	–	.0410	.8200	–	*.0690*	.8080	–	*.0290*	.8980	–	***.0540***	.7430	–
IFA	**.0296**	**.9337**	**.9592**	**.0348**	**.8613**	**.9356**	.0728	**.8206**	*.8812*	**.0277**	**.9175**	**.9587**	**.0470**	**.8057**	**.8942**

Fb is a commonly used evaluation metric in the field of information retrieval. In the task of salient object detection, following previous works [26], β^2 is usually set at 0.3. In this paper, we follow the parameters used in the previous studies.

4.2. Experimental comparison and analysis

4.2.1. Qualitative evaluation

The proposed method is compared with some state-of-the-art salient object detection methods, such as Amulet [27], RMMDF [25], GCPANet [10], PiCANet-R [7], PoolNet+ [9], R3Net [8], Bicon-P [2], SODGAN [12], DNA [28], GateNet [11], CANet [14], ITSD [13], A2S-v2 [26], AccoNet [29] and FPSI [30]. And the saliency maps of these methods are provided by the authors.

Fig. 6 shows the subjective results of our algorithm and the comparison methods. Image-1 shows the detection results containing complex background, Image-2 verifies the detection ability of the model in two objects with different sizes and Image-3 shows the effect of the model in objects with different depth. As we can see, our method can effectively highlight the boundary of the object while avoiding the amplification of noise, and can be applied to single object and multi-object salient detection. For image-3 in Fig. 6, the proposed method preserves the long-distance dolphin object, while in contrast the algorithms such as SODGAN and Bicon-P lose the object, and many of remaining methods in comparison produce blurred area.

4.2.2. Quantitative evaluation

Table 1 uses three metrics to verify the effectiveness of proposed method, and it can be seen that the proposed method (IFA) achieved the best performance in four datasets in terms of the listed metrics. In addition to that, the ITSD method also demonstrated excellent performance. On the Pascal-S dataset, the IFA method achieved the forth performance

Table 2
Performance comparison between proposed method and the other methods on RGB-D datasets: STERE and SIP.

	STERE			SIP		
$Methods_{Year}$	MAE	Fb	m-E	MAE	Fb	m-E
$HAINet_{21}$ [32]	.0393	.8837	.9332	.0534	.8732	.9149
$D3Net_{21}$ [17]	.0458	.8589	.9153	.0657	.8332	.8911
$Scribble_{22}$ [33]	.0485	.8516	.9238	.0626	.8264	.9050
$C2DFNet_{22}$ [34]	**.0372**	.8807	.9354	.0524	.8652	.9123
$MIRV_{22}$ [35]	.0413	.8710	.9318	.0497	.8613	.9227
$LSNet_{23}$ [36]	.0535	.8501	.9075	.0497	**.8813**	.9205
IFA	**.0372**	**.8933**	**.9406**	**.0480**	.8761	**.9231**

in terms of the MAE metric, and IFA has a small gap with SODGAN on mE (0.8823 vs 0.8812).

Table 2 presents a comprehensive comparative analysis between our proposed approach and state-of-the-art RGBD-based methodologies that leverage depth maps for the task of salient object detection. We carefully evaluated our model's performance on two meticulously chosen datasets: STERE and SIP. Our method performed outstandingly in most metrics of the two datasets. Among them, in the MAE of STERE, the results of our method were consistent with those of C2DFNet. In the Fb of SIP, our method ranked second with a minor disadvantage compared with LSNet (0.8761 vs 0.8813).

4.2.3. Subjective difference

To better show the effectiveness of different methods, we compare the generated salient maps separately. Specifically, we compare the pixel-wise difference between the results and the labels and save the pixel values (Image subtraction) as a two-dimensional numerical matrix. Finally, we obtain a difference map to evaluate the accuracy of different methods. As shown in Fig. 7, we convert the results into a colormap using the *ColorCube* space. The density of color in the difference map indicates the magnitude of the difference between the result and the label, with denser and richer colors indicating greater differences. As shown in Fig. 7, our method achieves the smallest difference from the label, while SODGAN also performs well. However, in the third image, the object detection with different depth information is incomplete.

In addition, we found that many comparison methods lose distant object (the second dolphin) in the image of the third row, such as the SODGAN method, while for the Amulet method, blurring occurs in edge-dense areas.

4.3. Ablation experiments and analysis

4.3.1. Effect of various structures

We considered ablation, one of which replaced the DFF block with a two-layer convolution (w/o DFF), and the other deleted the skip fusion part of the *DSFA* for features of different sizes (w/o DSFA). The role of our DFF module is to make features of different scales focus on their attention regions. Therefore, we set up a new experiment called $IFA(SE_{DFF})$, by replacing DFF with SE attention module. On the other hand, we analyzed the influence of different backbone on the results, including the VGG, ResNet and Swin-Transformer. The results are shown in Table 3. We observe that proposed method perform best on two datasets when we use Full model or ResNet50. Compared with different backbone such as Swin-T, VGG and ResNet101, the performance of IFA has increased by 64.9%, 22.5% and 29% on MAE respectively.

In order to explore the roles of *DFF* and *DSFA* in image feature extraction for SOD, we use feature visualization methods to demonstrate the effectiveness of the model. As shown in Fig. 8, the top three images are the results of replacing the *DFF* or *DSFA* block with the corresponding convolutional neural network. The structure of the bottom half is

Fig. 7. Examples of difference maps between the results of different methods and Ground truth. The difference maps are converted into the colormaps with the colrocube color space. Obviously, compared to other methods, our result is capable of not only locating salient objects but also refining the details of the detected salient objects. The difference between our results and the label is minimal.

Table 3
Performance comparison among ablation methods. "w/o" stands for "without", indicating the partial removal of the ablation module in Fig. 8(b) and (c). The lower part of the table represents the use of different backbones.

Methods	ECSSD			HKU-IS		
	MAE	Fb	m-E	MAE	Fb	m-E
w/o FR	.0308	.9195	.9541	.0377	.8371	.9198
w/o DFF	.0518	.8917	.9221	.0461	.8804	.9281
w/o DSFA	.0504	.8951	.9247	.0455	.8815	.9277
IFA(SE_{DFF})	.0327	.9191	.9033	.033	.8988	.9296
IFA(Swin-T)	.0843	.8203	.8649	.0719	.8056	.8768
IFA(VGG16)	.0382	.9075	.9437	.0343	.8921	.9446
IFA(ResNet-101)	.0417	.9084	.9393	.0391	.8911	.9393
IFA	**.0296**	**.9337**	**.9592**	**.0277**	**.9175**	**.9587**

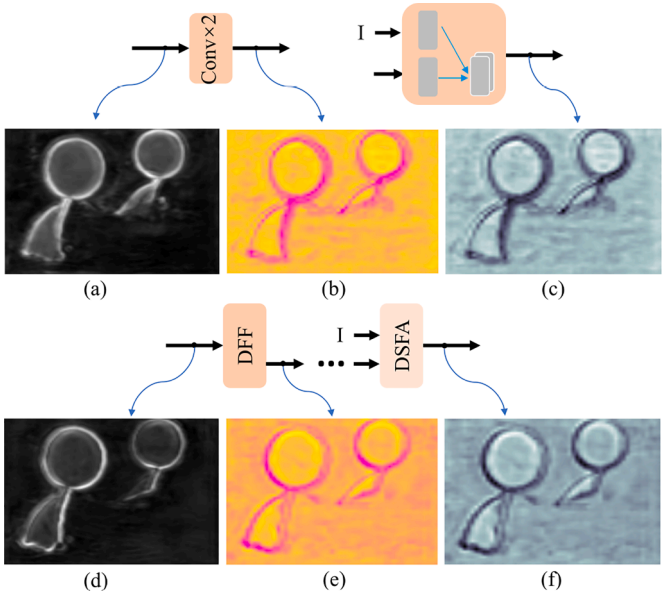


Fig. 8. Visualization of feature maps outputted by the ablation blocks. (a) and (b) Are from the model replacing DFF with convolution layers, and (d) and (e) Are from the model with DFF block. (c) Shows the feature map which is from the model without DSFA, (f) is from the proposed method.

the proposed method. We can see that the results processed by proposed method can eliminate certain artifacts compared to simple convolutional layers, and bring clearer edges. In particular, in Table 3, "FR" denotes the Feature Reduction module. Through the incorporation of FR, the model's accuracy and performance are notably improved.

4.3.2. *Effect of various inputs*

Here, we discuss the effects of different auxiliary inputs on the results. We choose two other approaches, namely Depth-based and Edge-based methods. For the Depth-based method, we use the NLPR dataset for comparison, which is an depth-based dataset. For the Edge-based method, we use the edge detection method Canny to obtain the edge maps of RGB images before training. The results are shown in Table 4 and NLPR-D is trained by STRTE dataset with depth maps and tested by NLPR and SSD dataset (these datasets are shown by Zhou et al. [20]). NLPR-I is tested by the same NLPR and SSD dataset. Specifically, we directly replace the light component with an edge map or a depth map for comparison.

In Table 4, "Edge" refers to the results obtained by using the Canny method to generate the edge estimation maps, using these instead of the illumination maps for both training and testing. "NLPR-D" represents

Table 4
Performance comparison with different auxiliary input. All the compared methods employ our proposed model.

Methods	MAE	Fb	m-E	MAE	Fb	m-E
	ECSSD			HKU-IS		
Edge	.0417	.9073	.9377	.0434	.8858	.9300
IFA	**.0296**	**.9337**	**.9592**	**.0277**	**.9175**	**.9587**
	NLPR			SSD		
NLPR-D	.0388	.8206	.9029	.0610	.7977	.8842
NLPR-I	**.0387**	**.8503**	**.9182**	**.0514**	**.8214**	**.9114**

the model trained using the NLPR dataset (with depth maps), while "NLPR-I" represents the model trained using the NLPR dataset (with generated illumination maps). As depth map labels are necessary, we selected the NLPR and SSD test sets for comparison.

4.4. *Model effectiveness*

In order to fully evaluate the performance of the model, we set up experiments on the efficiency of the model. According to the existing methods, most of the current methods [17] are based on pre-trained backbone to extract image features. Deeper features often contain larger feature channels, which will lead to a large number of parameters in the model. To solve this problem as much as possible, we use a specially designed FR module, which is composed of a series of four convolution layers. It is mainly used to gradually reduce the number of channels in the model and reduce the model parameters.

Table 5 shows the comparison experiments of our method with some excellent methods in terms of model parameters and frames per second (FPS). And we analyze the effects of various inputs on the model, including single RGB images and RGBD-based methods. Our method occupies 54.74M parameters, and in real-time processing, the FPS of our model can reach 26.86 frames per second (input image size: 256*256).

For the single-input model, our method outperforms most methods in model parameters and FPS, while inferior to methods Amulet and PiCANet in model parameters. Despite the advantages of the above two methods in model parameters, our method outperforms them in FPS.

For RGBD-based methods, except for C2DFNet and D3Net, our method has advantages in terms of both model parameters and FPS performance. Compared with C2DFNet, IFA has 3M more parameters in the model parameters, but IFA has an improvement in FPS. For D3Net [17], it is better than IFA both in terms of model parameters and FPS. However, it is worth noting that although it achieves a smaller model size,

Table 5
Performance comparison with different methods on model effectivity.

$Methods_{Pub/Year}$	Input[3] (Type)	Param (M)	FPS (256x256)
$Amulet_{ICCV17}$ [27]	Single	33.16	20[1]
$R3Net_{IJCAI18}$ [8]	Single	70.18	–
$PiCANet_{TIP18}$ [7]	Single	47.22	7
$PoolNet_{TPAMI22}$ [9]	Single	68.26	–
$TriTransNet_{MM21}$ [19]	Dual(D)	139.55	8.62
$D3Net_{TNNLS21}$ [17]	Dual(D)	45.24[2]	29.85
$C2DFNet_{TMM22}$ [34]	Dual(D)	51.61	26.78
$MIRV_{TCSVT23}$ [35]	Dual(D)	149.34	–
IFA(w/o FR)	Dual(I)	256.13	10.66
IFA	Dual(I)	54.74	26.86

[1] The data are from Ref [9].
[2] D3Net contains three same branch models.
[3] ***Single*** represents a single RGB image as input, and ***Dual*** represents a dual input model with auxiliary features, where ***D*** and ***I*** represent the use of depth features and the illumination maps.

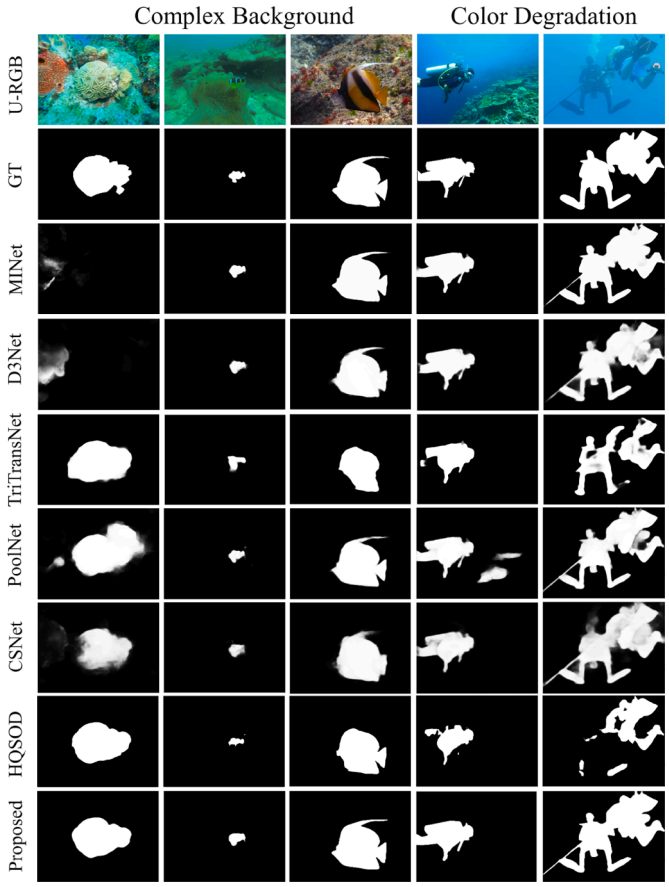


Fig. 9. A few qualitative comparisons of salient maps generated by different methods on USOD dataset. The first three columns show images of a complex underwater background, and the last two columns show images of fading underwater light.

it is worth noting that although it achieves a smaller model size, it is inferior to us in the objective evaluation metrics we list (See in Table 2).

5. Expended experiments

5.1. Expended experiments on low-level scenes

5.1.1. Underwater-based results and analysis

Just like the task on land, underwater SOD is also a meaningful study. The ocean is filled with abundant resources, and deep-sea exploration is increasingly becoming a mainstream technological field. However, due to complex backgrounds, scattering of seawater, and other factors, underwater objects are more difficult to identify, and salient objects are harder to distinguish. We selected the existing underwater SOD dataset (USOD10K [37]) and compared it with state-of-the-art methods. Fig. 9 shows the results of subjective comaprison among some advances methods on USOD10K, which includes 1026 underwater images.

Fig. 9 displays a visual comparison of multiple methods on the USOD dataset. It includes underwater images with complex backgrounds and also showcases severely color degraded images. In the first column of images, our method is able to better restore the contours of the objects. In the third column, the HQSOD [37] and TriTrans methods failed to detect the presence of fish fins. For the severely color degraded images in the fifth column, our method consistently outperforms the listed methods (by comparing the completeness of the diver's contour on the left side). We have selected the six recent state-of-the-art methods to compare their objective performance. As shown in Table 6, our method outperforms the listed methods in all five evaluation metrics (among them, the best performance is highlighted in bold).

Table 6
Objective performance of different methods on USOD10K.

Methods	Pub/Year	Fb	max-F	MAE	max-E	m-E
MINet [38]	$CVPR_{20}$	.8855	.9072	.0287	.9501	.9393
D3Net [17]	$TNNLS_{21}$	.8408	.8807	.0374	.9413	.9107
TriTrans [19]	MM_{21}	.7466	.7501	.0659	.8479	.8327
HQSOD [37]	$ICCV_{21}$	.7677	.7714	.0552	.8388	.8267
SVAM-Net [39]	RSS_{22}	.6251	.6451	.0915	.7649	.7466
CSNet [40]	$TPAMI_{22}$	.7825	.8462	.0548	.9178	.8652
IFA	–	**.8892**	**.9098**	**.0242**	**.9582**	**.9527**

5.1.2. Illumination-aware results and analysis

For salient object detection tasks, acquiring and annotating real low-light scene images is a challenging task. Consequently, we primarily validate the effectiveness of our method through two main approaches: subjective evaluation of individual images and objective evaluation of the Synthetic dataset.

Subjective evaluation of individual images. Real world scenes contain more low-quality environments (uneven lighting, complex background, etc.), which also brings great challenges to the task of salient object detection. The existing data sets also contain a large number of factors such as uneven illumination and shadows. Fig. 10(a) shows the proportion of relatively low illumination images contained in the five known data sets. We use the pixel distribution of the image (convert to grayscale) to measure the brightness of the image, and take the thresholds of 50, 60, 70, 80 and 100 according to the average value of pixel. Taking PASCAL-S dataset as an example, Fig. 10(b) shows the test results of the data set at average pixel less than 80. IFA can effectively improve the performance of sod detection in low illumination environment.

Low illumination image is not only difficult for people to find salient objects, but also affects the accuracy of computer recognition. By introducing Retinex theory, salient object detection can be effectively completed in the scene of low illumination. Fig. 11 shows the comparison of the detection result of salient objects contain more noise and shadow, which is not conducive to the segmentation of clear boundaries in the model. However, our method can eliminate the influence of illumination.

Objective evaluation of the Synthetic dataset. Objective evaluation metrics can better measure the performance of different methods in low-light scenarios. However, existing datasets have not been effectively applied to low-light salient object detection tasks while maintaining the accuracy of image labels. To address this dataset issue while preserving the accuracy of image labels, we generated low-light images by applying a pixel-wise exponential transformation function. In this study, We selected all low-light images from the ECSSD and DUTS-TE datasets (using an average pixel value threshold of 60) to create two new test sets, LECSSD and LDUTS-TE. We conducted tests on these two datasets to compare their objective performance. As shown in Table 7, our method

Table 7
Objective performance comparison on LECSSD and LDUTS-TE.

Methods	MAE	Fb	m-E	MAE	Fb	m-E
	LECSSD			LDUTS-TE		
Amulet [27]	.0681	.8773	.9020	.0792	.7271	.8339
AFNet [31]	.0347	.9281	.9514	.0549	.8035	.8831
PiCANet-R [7]	.0458	.8952	.9294	.0633	.7846	.8637
PoolNet [9]	.0407	.9223	.9455	.0447	.8264	.8877
Bicon-P [2]	.0433	.9118	.9283	.0559	.8030	.8682
R3Net [8]	.0639	.8966	.9080	.0676	.7767	.8579
SODGAN [12]	.0476	.8958	.9328	.0394	.8142	.9154
GCPANet [10]	.0385	.8992	.9267	.0399	.8476	.9076
GateNet [11]	.0384	.9219	.9377	.0509	.8186	.8792
DNA [28]	.0524	.8886	.9232	.0476	.8031	.8960
ITSD [13]	**.0319**	.9276	.9524	.0457	.8443	.9099
A2S-v2 [26]	.0507	.9172	.9330	.0524	.8351	.9027
IFA	**.0319**	**.9282**	**.9591**	**.0334**	**.8945**	**.9480**

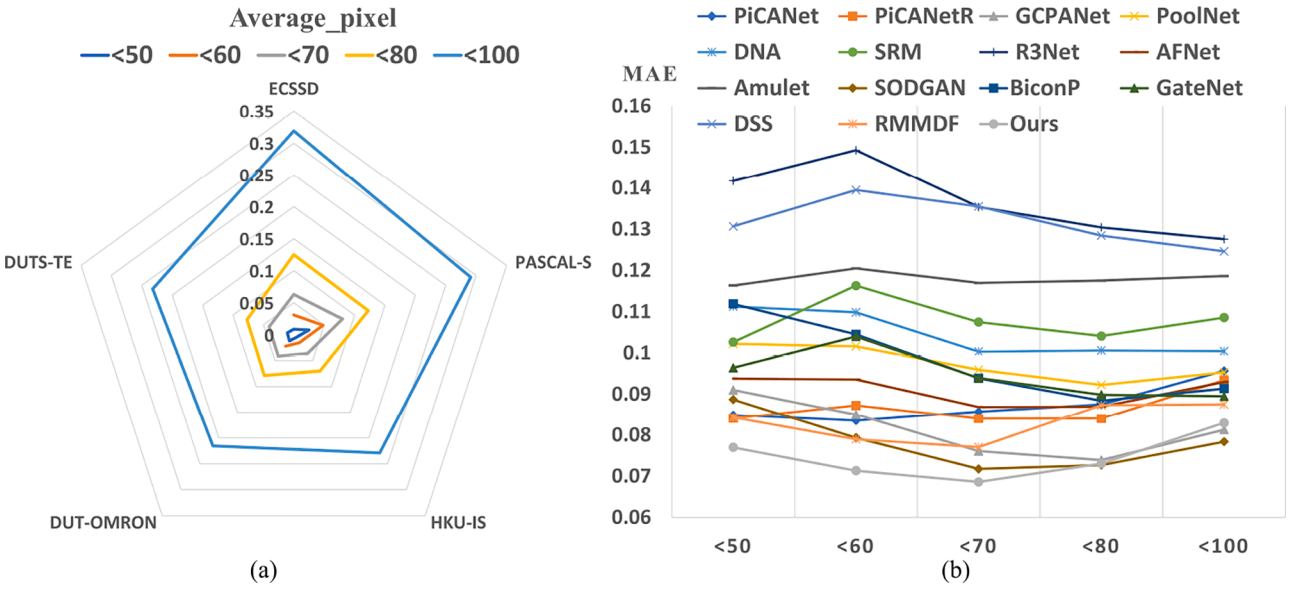


Fig. 10. Salient object detection results in low light images with different pixel threshold. (a) Represents the proportion of low-light images complying with different thresholds based on pixel average values in the five datasets. (b) Illustrates the MAE (Mean Absolute Error) results of our method and the compared methods at different thresholds on the PASCAL-IS dataset.

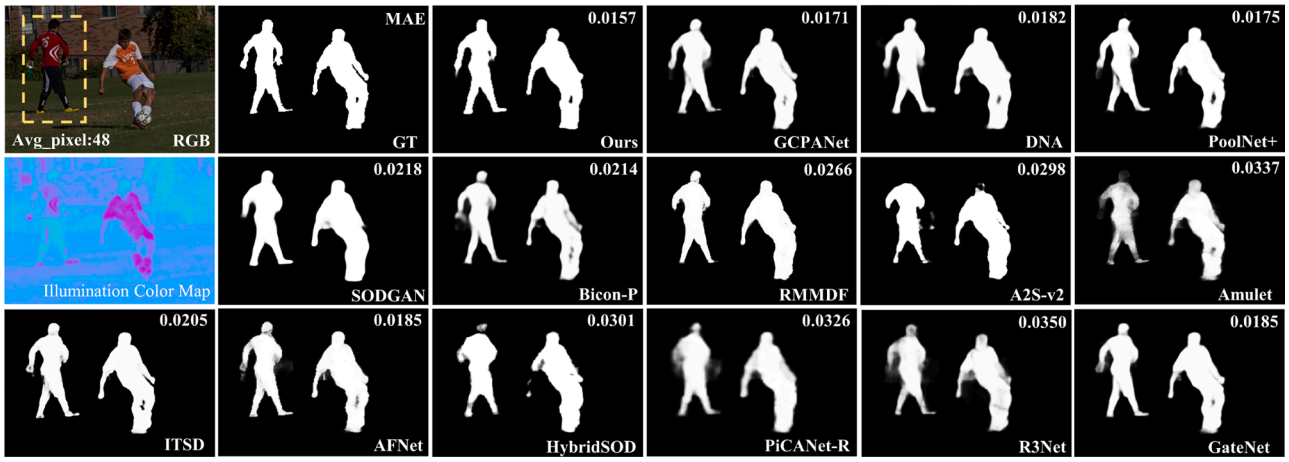


Fig. 11. An example of SOD in low-light image (from the ECSSD Dataset). The text in bottom-right corner is represents the method and the digit in the top-right corner is shows the result of MAE.

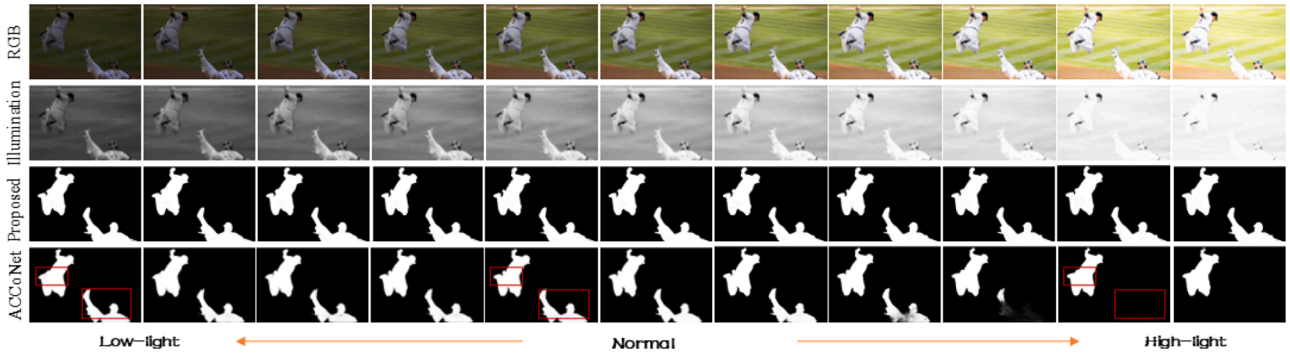


Fig. 12. Comparison of visual results in different light scenes. We can find that our method can maintain relatively consistent results in low/high-light scenes.

achieved excellent performance on LECSSD and LDUTS-TE dataset, effectively identifying salient objects in low-light scenes, thus meeting the requirements of the task.

In addition, we use multi-exposure scenes to verify our effect. Fig. 12 shows the visual comparison between our method and ACCoNet under scenes with different exposure rates. We can find that our method can perceive illumination information, and the segmentation results are rarely affected by strong or weak light, while the comparison method often show inconsistent results in low/high light environments.

6. Conclusion and future works

This paper explores the potential of the illumination maps by introducing the Retinex theory on salient object detection. In this study, the illumination map of the image is obtained by using the decomposition network (*Decom-Net*), and then the illumination map is fed into the proposed model together with the RGB image. Then we design a Diverse Feature Fusion Block (*DFF*) and Up-Sample Feature Extension Block (*DSFE*) to extract the features of RGB image. At the same time, we designed Down-Sample Feature Aggregation Block (*DSFA*) to better integrate the illumination maps, and the blocks are fused with features of different scales. The major novelty of the study lies in the discovery of the illumination maps obtained by Retinex theory that can be used for SOD well, which creates a new idea for SOD. In the future, we will focus on exploring the feature processing tasks of illumination maps, while extending the task to video-based researches. For illumination map, it still contains a large amount of redundancy under complex background. In the subsequent work, we will further extract salient regions based on illumination processing. According to the research of Yang et al. [22] and Jing et al. [21], we can next explore the specific role of different modules in the model for sod tasks, so as to effectively realize the efficiency of the model.

CRediT authorship contribution statement

Miao Li: Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Formal analysis, Conceptualization; **Hongyun Zhang:** Writing – review & editing, Supervision, Funding acquisition; **Kecan Cai:** Writing – review & editing; **Witold Pedrycz:** Writing – review & editing, Conceptualization; **Duoqian Miao:** Writing – review & editing, Supervision, Funding acquisition; **Ying Gao:** Writing – review & editing.

Data availability

Data will be made available on request.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grants 62076182, 62376198, 62376199 and 62076184. The National Key Research and Development Program of China under Grants 2022YFB3104700. The Natural Science Foundation of Shanghai under Grant 22ZR1466700.

References

[1] Y. Zheng, J. Yang, H. Tao, Y. Wang, L. Chen, Y. Wang, T. Cao, Self-distillation salient object detection via generalized diversity loss, Pattern Recogn. 168 (2025) 111804. https://doi.org/10.1016/j.patcog.2025.111804
[2] Z. Yang, S. Soltanian-Zadeh, S. Farsiu, BiconNet: an edge-preserved connectivity-based approach for salient object detection, Pattern Recogn. 121 (2022) 108231.
[3] H. Zhu, J. Ni, X. Yang, L. Zhang, CMIGNet: cross-modal inverse guidance network for RGB-depth salient object detection, Pattern Recogn. 155 (2024) 110693.
[4] J. Wang, Z. Zhang, N. Yu, Y. Han, Progressive expansion for semi-supervised bi-modal salient object detection, Pattern Recogn. 157 (2025) 110868.
[5] R. Cong, K. Zhang, C. Zhang, F. Zheng, Y. Zhao, Q. Huang, S. Kwong, Does thermal really always matter for RGB-T salient object detection?, IEEE Trans. Multimed. 25 (2023) 6971–6982. https://doi.org/10.1109/TMM.2022.3216476
[6] M. Li, D. Zhou, R. Nie, S. Xie, Y. Liu, AMBCR: low-light image enhancement via attention guided multi-branch construction and Retinex theory, IET Image Process. 15 (9) (2021) 2020–2038.
[7] N. Liu, J. Han, M.H. Yang, PiCANet: pixel-wise contextual attention learning for accurate saliency detection, IEEE Trans. Image Process. 29 (2020) 6438–6451.
[8] Z. Deng, X. Hu, L. Zhu, X. Xu, J. Qin, G. Han, P.-A. Heng, R3net: recurrent residual refinement network for saliency detection, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence, AAAI Press Menlo Park, CA, USA, 2018, pp. 684–690.
[9] J.J. Liu, Q. Hou, Z.A. Liu, M.M. Cheng, Poolnet+: exploring the potential of pooling for salient object detection, IEEE Trans. Pattern Anal. Mach. Intell. 45 (1) (2022) 887–904.
[10] Z. Chen, Q. Xu, R. Cong, Q. Huang, Global context-aware progressive aggregation network for salient object detection, in: Proceedings of the 27th International Joint Conference on Artificial Intelligence 34, 2020, pp. 10599–10606.
[11] X. Zhao, Y. Pang, L. Zhang, H. Lu, L. Zhang, Suppress and balance: a simple gated network for salient object detection, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16, Springer, 2020, pp. 35–51.
[12] Z. Wu, L. Wang, W. Wang, T. Shi, C. Chen, A. Hao, S. Li, Synthetic data supervised salient object detection, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 5557–5565.
[13] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, L. Yang, Interactive two-stream decoder for accurate and fast saliency detection, in: IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020.
[14] J. Li, Z. Pan, Q. Liu, Y. Cui, Y. Sun, Complementarity-aware attention network for salient object detection, IEEE Trans. Cybern. 52 (2) (2022) 873–886. https://doi.org/10.1109/TCYB.2020.2988093
[15] Y. Wang, R. Wang, X. He, C. Lin, T. Wang, Q. Jia, X. Fan, WBNet: weakly-supervised salient object detection via scribble and pseudo-background priors, Pattern Recogn. 154 (2024) 110579.
[16] Y. Yi, N. Zhang, W. Zhou, Y. Shi, G. Xie, J. Wang, GPONet: a two-stream gated progressive optimization network for salient object detection, Pattern Recogn. 150 (2024) 110330.
[17] D. Fan, Z. Lin, Z. Zhang, M. Zhu, M. Cheng, Rethinking RGB-D salient object detection: models, data sets, and large-scale benchmarks, IEEE Trans. Neural Netw. Learn. Syst. 32 (5) (2021) 2075–2089.
[18] L. Chen, D. Zhang, X. Wang, C. Wan, S. Jin, Z. Zheng, A complementary dual model for weakly supervised salient object detection, Pattern Recogn. 163 (2025) 111465.
[19] Z. Liu, Y. Wang, Z. Tu, Y. Xiao, B. Tang, TriTransNet: RGB-D salient object detection with a triplet transformer embedding network, in: Proceedings of the 29th ACM International Conference on Multimedia, 2021, pp. 4481–4490.
[20] T. Zhou, D.P. Fan, M.M. Cheng, J. Shen, L. Shao, RGB-D salient object detection: a survey, Comput. Visual Media 7 (1) (2021) 37–69. https://doi.org/10.1007/s41095-020-0199-z
[21] Y. Jing, C. Yuan, L. Ju, Y. Yang, X. Wang, D. Tao, Deep graph reprogramming, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 24345–24354.
[22] X. Yang, D. Zhou, S. Liu, J. Ye, X. Wang, Deep model reassembly, Adv. Neural Inform. Process. Syst. 35 (2022a) 25739–25753.
[23] X. Yang, J. Ye, X. Wang, Factorizing knowledge in neural networks, In European Conference on Computer Vision (2022b).
[24] M. Li, L. Zhao, D. Zhou, R. Nie, Y. Liu, Y. Wei, AEMS: an attention enhancement network of modules stacking for lowlight image enhancement, Vis. Comput. 38 (12) (2022) 4203–4219.
[25] Z. Wu, S. Li, C. Chen, A. Hao, H. Qin, Recursive multi-model complementary deep fusion for robust salient object detection via parallel sub-networks, Pattern Recogn. 121 (2022) 108212.
[26] H. Zhou, B. Qiao, L. Yang, J. Lai, X. Xie, Texture-guided saliency distilling for unsupervised salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 7257–7267.
[27] P. Zhang, D. Wang, H. Lu, H. Wang, X. Ruan, Amulet: aggregating multi-level convolutional features for salient object detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2017, pp. 202–211.
[28] Y. Liu, M.M. Cheng, X.Y. Zhang, G.Y. Nie, M. Wang, DNA: deeply supervised nonlinear aggregation for salient object detection, IEEE Trans. Cybern. 52 (7) (2021) 6131–6142.
[29] G. Li, Z. Liu, D. Zeng, W. Lin, H. Ling, Adjacent context coordination network for salient object detection in optical remote sensing images, IEEE Trans. Cybern. 53 (1) (2023) 526–538. https://doi.org/10.1109/TCYB.2022.3162945
[30] X. Wang, Z. Liu, L. Veronica, Z. Huang, Feature specific progressive improvement for salient object detection, Pattern Recogn. 147 (2024) 110085. https://doi.org/10.1016/J.PATCOG.2023.110085
[31] M. Feng, H. Lu, E. Ding, Attentive feedback network for boundary-aware salient object detection, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019, pp. 1623–1632.
[32] G. Li, Z. Liu, M. Chen, Z. Bai, W. Lin, H. Ling, Hierarchical alternate interaction network for RGB-D salient object detection, IEEE Trans. Image Process. 30 (2021) 3528–3542.
[33] Y. Xu, X. Yu, J. Zhang, L. Zhu, D. Wang, Weakly supervised RGB-D salient object detection with prediction consistency training and active scribble boosting, IEEE Trans. Image Process. 31 (2022) 2148–2161. https://doi.org/10.1109/TIP.2022.3151999
[34] M. Zhang, S. Yao, B. Hu, Y. Piao, W. Ji, C2DFNet: criss-cross dynamic filter network for RGB-D salient object detection, IEEE Trans. Multimed. 25 (2022) 1–13.
[35] A. Li, Y. Mao, J. Zhang, Y. Dai, Mutual information regularization for weakly-supervised RGB-D salient object detection, IEEE Trans. Circ. Syst. Video Technol. (2023) 397–410. https://doi.org/10.1109/TCSVT.2023.3285249

[36] W. Zhou, Y. Zhu, J. Lei, R. Yang, L. Yu, LSNet: lightweight spatial boosting network for detecting salient objects in RGB-thermal images, IEEE Trans. Image Process. 32 (2023) 1329–1340. https://doi.org/10.1109/TIP.2023.3242775
[37] L. Bo, T. Lv, Y. Zhong, S. Ding, M. Song, Disentangled high quality salient object detection, in: 2021 IEEE/CVF International Conference on Computer Vision, 2021, pp. 3560–3570.
[38] Y. Pang, X. Zhao, L. Zhang, H. Lu, Multi-scale interactive network for salient object detection, in: 2020 IEEE Conference on Computer Vision and Pattern Recognition, 2020, pp. 9410–9419.
[39] M.J. Islam, R. Wang, J. Sattar, SVAM: saliency-guided visual attention modeling by autonomous underwater robots, in: Robotics: Science and Systems (RSS), NY, USA, 2022.
[40] M.M. Cheng, S.H. Gao, A. Borji, Y.Q. Tan, Z. Lin, M. Wang, A highly efficient model to study the semantics of salient object detection, IEEE Trans. Pattern Anal. Mach. Intell. 44 (11) (2022) 8006–8021.

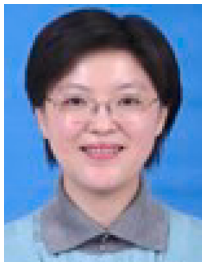
Miao Li is currently pursuing his Ph.D. in Computer Sciences at Tongji University, Shanghai, China. He received his master's degree from School of Information science and Engineering, Yunnan University in 2022. His research interests include computer vision and pattern recognition, image enhancement, object detection and image segmentation. E-mail:lmiao@tongji.edu.cn.

Hongyun Zhang received the Ph.D. degree in pattern recognition and intelligence system from Tongji University, Shanghai, China, in 2005. She is doctoral supervisor and currently an Associate Professor at Tongji University. She is the author or co-author of nearly 70 journal papers and conference proceedings in principal curves, pattern recognition, machine learning granular computing, and rough set Her current research interests include computer vision and pattern recognition, principal curves, data mining, rough set theory, and granular computing. E-mail: zhanghongyun@tongji.edu.cn.

Kecan Cai is currently pursuing his Ph.D. in Computer Sciences at Tongji University, Shanghai, China. His research interests include image classification, pattern recognition and image segmentation. E-mail: caikecan@tongji.edu.cn.

Witold Pedrycz received the M.Sc. degree in computer science and technology, the Ph.D. degree in computer engineering, and the D.Sci. degree in system science from the Silesian University of Technology, Gliwice, Poland, in 1977, 1980, and 1984, respectively. He is a Professor with the Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada, the Department of Electrical and Computer Engineering, Faculty of Engineering, King Abdulaziz University, Jeddah, Saudi Arabia, and the Systems Research Institute, Polish Academy of Sciences, Warsaw, Poland. He has authored 17 research monographs covering various aspects of computational intelligence, data mining, and software engineering. He is an Editor-in-Chief of Information Sciences, WIREs Data Mining and Knowledge Discovery (Wiley), and the International Journal of Granular Computing (Springer). He is a Fellow of IEEE and the Royal Society of Canada and a Foreign Member of the Polish Academy of Sciences. E-mail: wpedrycz@ualberta.ca.

Duoqian Miao is professor of College of Electronics and Information Engineering of Tongji University, Fellow of International Rough Set Society (IRSS), Fellow of Chinese Association for Artificial Intelligence (CAAI). Prof. Miao works in Department of Computer Science and Technology of Tongji University. Prof. Miao's research interests include Artificial Intelligence, Machine Learning, Big Data Analysis, Granular Computing and Rough Sets, etc. He has published more than 160 papers in this area, more than nine books and academic works, and nine national invention patents. E-mail: dqmiao@tongji.edu.cn.



Ying Gao is professor of School of Management and Engineering of Capital Univetsity of Economic and Business, Beijing, China. She received the Ph.D. degree from Renmin University of China, Beijing, in 2006. Her current research interests include computer vision, data mining and information management. E-mail: gaoying517@cueb.edu.cn.