RGB-D visual saliency detection of stacked fruits under poor lighting

Chunjian Hua[1,2*], Xintong Zou[1,2], Yi Jiang[1,2], Jianfeng Yu[1,2], Ying Chen[3]

(1. *School of Mechanical Engineering, Jiangnan University, Wuxi 214122, Jiangsu, China*;
2. *Jiangsu Key Laboratory of Advanced Food Manufacturing Equipment & Technology, Jiangnan University, Wuxi 214122, Jiangsu, China*;
3. *School of Internet of Things Engineer, Jiangnan University, Wuxi 214122, Jiangsu, China*)

Abstract: The saliency detection of the same kind of stacked fruits can assist robots in completing sorting tasks, which is an important prerequisite for the grading and packing of fruits. In order to accurately obtain saliency targets of fruits in the same kind of stacked state under overexposure, non-uniform illumination, and low illumination, a method for detecting stacked fruits under poor illumination based on RGB-D visual saliency was proposed. Based on the Res2Net network, features from each layer of two images were obtained. To realize the complementary advantages between RGB features and depth features, the input RGB images were preprocessed using depth weighting to obtain purified RGB features. To increase the information interaction between branches of different scales and better balance the fusion features and modal exclusive features, a multi-scale progressive fusion module was proposed. To minimize the difference between the initial saliency maps generated by different features and improve the accuracy of the final predicted saliency maps, a multi-branch hybrid supervised method was used. The comprehensive experiments on the self-made dataset of the same kind of stacked fruits show that the proposed algorithm is superior to five state-of-the-art RGB-D SOD methods in four key indicators: *S* value, *F* value, and MAE value, which are 0.979, 0.992, and 0.006, respectively, and the *P-R* curve, which is also closer to the upper right corner of the graph. These values demonstrate that the proposed algorithm can accurately obtain saliency targets in the same kind of stacked fruits. The results of this study can promote the automatic development of the fruit production and packaging industry.

Keywords: RGB-D salient object detection, multi-branch fusion, depth weighting, mixed supervision, same kind of stacked fruits

DOI: 10.25165/j.ijabe.20251801.8057

Citation: Hua C J, Zou X T, Jiang Y, Yu J F, Chen Y. RGB-D visual saliency detection of stacked fruits under poor lighting. Int J Agric & Biol Eng, 2025; 18(1): 230–237.

1 Introduction

Fruit sorting is an important link to realizing the intelligence of the fruit industry chain[1], and it is also the primary prerequisite for fruit injury detection[2], grading, and packing. At present, fruit sorting in the same stacking state mainly depends on a large number of people or large equipment with vibration[3]. These methods have problems such as low manual efficiency, large equipment areas, and easy to cause fruit damage. Therefore, it is imperative to use robots to complete the sorting task. However, if you want the robot to grasp the top fruit directly like in manual operation, you must be able to detect the foreground salient targets in the same kind of stacked fruits.

Because the difference between foreground fruit target and background in the same stacking state is too small, it is difficult to segment by traditional visual methods. If you want to obtain the accurate foreground target to be captured, salient object detection (SOD)[4] is the preferred method. The goal of SOD is to detect and highlight the most salient objects in the visual input, which has been applied to many computer vision tasks, such as remote sensing images[5] and biomedicine[6], etc.

As the depth map complements the distance information, it can improve the detection accuracy under complex backgrounds to a certain extent. RGB-D SOD has attracted more and more research attention, and many RGB-D SOD methods have been proposed and made some advances. Arivazhagan et al.[7] generated saliency maps from RGB and depth action sequences, extracted symbols, amplitude, and center descriptors representing complete local binary patterns from them, and then fused depth features and RGB features through canonical correlation analysis and dimension reduction. Liu et al.[8] proposed a dual stream thinning network, designed a fusion thinning module to fuse the output features of different resolutions and models, and used low layer depth features with higher resolution to refine the boundary of detected targets. Zhao et al.[9] applied the contrast prior to the CNN-based architecture to enhance the depth information, and further integrated the enhanced depth features with RGB features using a new fluid pyramid integration. Singh et al.[10] proposed a composite backbone network with mutual attention-based discrimination windows. At each encoder stage, discrimination windows based on channel, spatial, and feature level attention are inserted to enhance salient features. Das et al.[11] used the depth estimation network to find the depth map of a two-dimensional image, and used the depth map to train the depth-guided saliency network to generate an intermediate depth saliency map. Finally, they fused the depth saliency maps with the rough saliency maps to obtain the final saliency maps. The main limitation of the above methods is that the influence of lighting conditions has not been taken into account. To simplify the volume of sorting equipment, if the factory lighting is directly used as the light source,

Received date: 2022-11-28 **Accepted date:** 2023-03-15

Biographies: **Xintong Zou**, MS, research interest: machine vision and deep learning, Email: 2486019375@qq.com; **Yi Jiang**, PhD, Associate Professor, research interest: Linux-based open CNC system and robot control system, Email: jiangyi@jiangnan.edu.cn; **Jianfeng Yu**, PhD, Professor, research interest: robot design and control, embedded equipment, Email: robotmcu@126.com; **Ying Chen**, PhD, Professor, research interest: computer vision, pattern recognition, and information fusion, Email: 10699785@qq.com.

***Corresponding author: Chunjian Hua**, PhD, Associate Professor, research interest: machine vision and pattern recognition. School of Mechanical Engineering, Jiangnan University, Wuxi 214122, Jiangsu, China. Tel: +86-13921189751, Email: 277795559@qq.com.

there will inevitably be overexposure, non-uniform illumination, and low illumination. Poor lighting conditions will inevitably affect the detection accuracy.

To address the above problems, a stacked fruits detection algorithm under poor lighting based on RGB-D visual saliency was proposed in this study. The main contributions of this method are as follows:

1) The depth-weighted preprocessing module was introduced to purify the input image pairs, retain the features of the input image pairs that are conducive to salient target detection, weaken the faculae and other misleading features, and avoid the impact of unstable image quality of the input images on the detection accuracy;

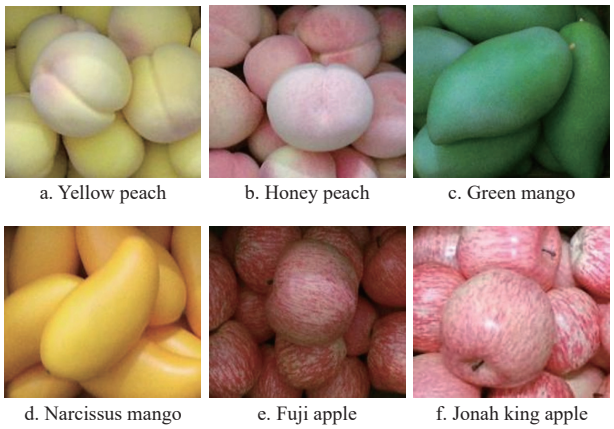
2) A multi-scale progressive fusion module was proposed, so that the proposed algorithm can effectively retain the exclusive features of modes in the process of multimodal feature fusion, increase the information interaction between branches of different scales, fully cover the context information, and maximize the utilization of input features, thereby weakening the influence of lighting conditions and improving the detection accuracy;

3) The hybrid supervision method was adopted for the initial saliency maps generated by multiple branches so that the pixel-level binary cross-entropy loss and the map-level intersection-over-union loss complement each other, accelerate algorithm convergence, and reduce the impact of adverse factors such as reflection and low illumination, to improve the accuracy of saliency prediction.

2 Materials and methods

2.1 Image acquisition

The datasets of this study consist of two public datasets and a self-made dataset. The public datasets include 1485 images from NJU2K[12] and 700 images from NLPR[13], and the self-made dataset includes 300 images, covering six kinds of fruits, namely yellow peach, honey peach, green mango, narcissus mango, Fuji apple, and Jonah king apple, as shown in Figure 1.



a. Yellow peach b. Honey peach c. Green mango

d. Narcissus mango e. Fuji apple f. Jonah king apple

Figure 1 Samples of six types of stacked fruits

To enhance the robustness of the network to image transformation and alleviate the problem of overfitting, data enhancement was used, and random rotation, random clipping, and horizontal flipping were applied to the images in the training dataset. In this study, 300 images of the self-made dataset were divided into the training set and test set according to a ratio of 8:2. Table 1 lists the number of images of various stacked fruits in the dataset.

Table 1 The number of images of various stacked fruits

Stacked fruit type	Graphics	Training set	Test set
Yellow peach	52	42	10
Honey peach	46	36	10
Green mango	59	49	10
Narcissus mango	48	38	10
Fuji apple	46	36	10
Jonah king apple	49	39	10

2.2 Network of the same kind of stacked fruit detection

As shown in Figure 2, the framework of the stacked fruits detection algorithm is based on RGB-D visual saliency under poor lighting. The backbone network is Res2Net-101[14], and the inputs are RGB and depth image pairs of the same kind of stacked fruits. The model structure is divided into three parts: first, depth-weighted preprocessing is performed on the input RGB images to obtain purified RGB features; then the fusion features are obtained by progressively fusing the purified RGB features and depth features of each layer; finally, in the decoding stage, the initial saliency maps generated by RGB branch, depth branch, and fusion branch are supervised by the multi-branch hybrid supervision method to improve the accuracy of the final prediction saliency maps.

2.2.1 Depth-weighted preprocessing

The unstable quality of the input image pairs is a key factor affecting the accuracy of RGB-D saliency detection. High-quality RGB and depth image pairs have the feature of "edge alignment"[15]. Therefore, this feature can be used to retain the features of the input image pair that are conducive to the detection of salient objects and reduce faculae and other misleading features, so as to improve the detection accuracy. Inspired by the DFM algorithm[15], the depth-weighted preprocessing (DWP) module in Figure 2 was introduced. The specific structure of the DWP module is shown in Figure 3.

To obtain rich edge features, the low-dimensional features $r1$ and $d1$ of RGB stream and depth stream in Figure 2 are first input into the DWP module, and then convolved[16] to obtain the edge activation features $r1'$ and $d1'$. The edge alignment feature vector Vec1 between $r1$ and $d1$ is shown in Equation (1).

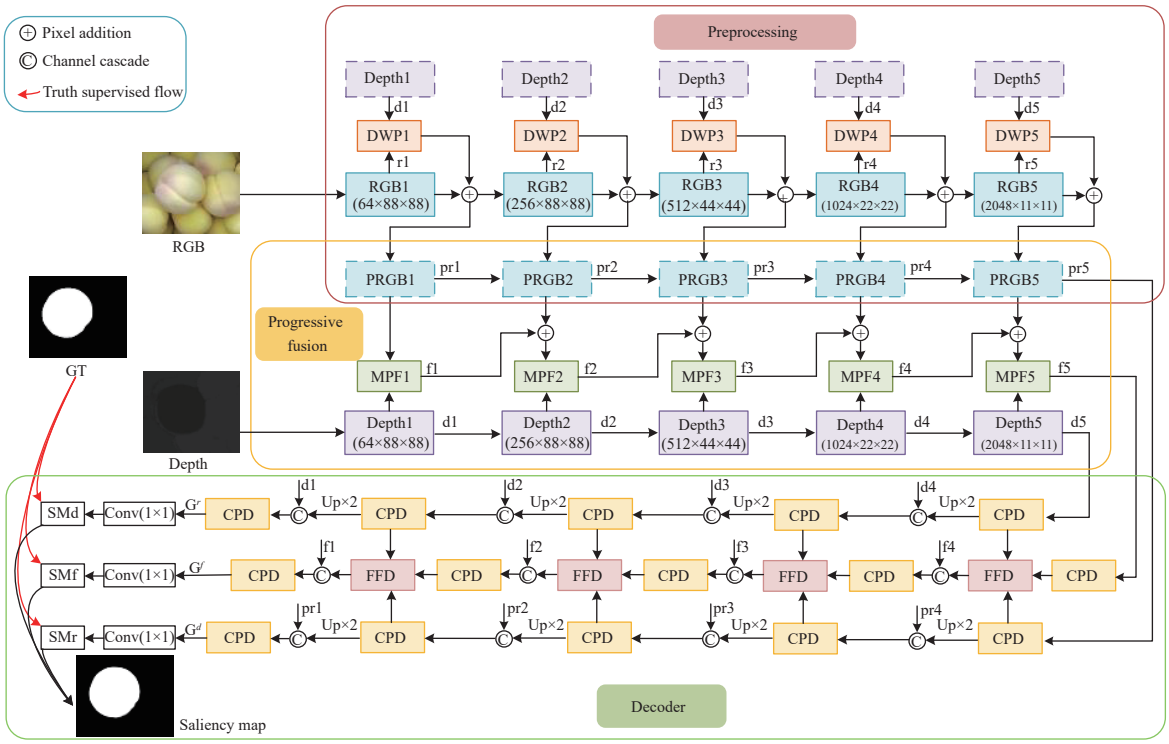
$$\mathrm{Vec1}=\frac{\mathrm{GAP}(r1'\otimes d1')}{\mathrm{GAP}(r1'\oplus d1')} \tag{1}$$

where, GAP represents the global average pooling operation.

To make the edge alignment feature vector robust to slight disturbances, $r1$ and $d1$ are sampled down twice by using the maximum pooling with a step of 2, and Vec1 is calculated on multiple scales. The calculation method is the same as that of Equation (1) to obtain Vec2 and Vec3. Vec1, Vec2, and Vec3 are cascaded to generate the enhancement vector Evec. The edge enhancement vector v_i (i=1, 2, 3, 4, 5) is obtained by separating the elements of the enhancement vector Evec.

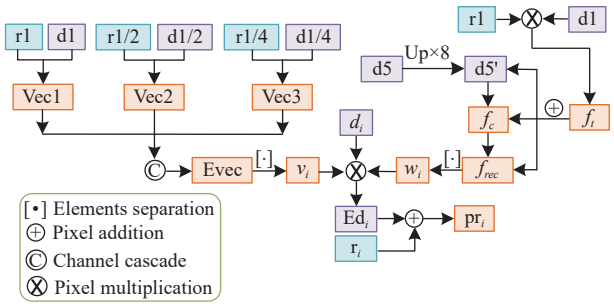
Because the highest level feature $d5$ from the depth stream can realize the coarse position of salient areas, the $d5$ is recalibrated through low-dimensional RGB and depth features to improve its position accuracy for salient areas.

First, an 8 times up-sampling operation is performed on $d5$ to make it the same size as $r1$ and $d1$, and $d5'$ in Figure 3 is obtained. Then, $r1$ and $d1$ are multiplied by elements to generate a common edge feature f_t. To better simulate the long-term correlation between low-dimensional features and high-dimensional features, the maximum pooling operation and dilated convolution are used to rapidly increase the receptive fields. The recalibration process is shown in Equation (2).



DWP: Depth-weighted preprocessing; PRGB: Preprocessed RGB; MPF: Multi-scale progressive fusion; CPD: Cascaded partial decoder; FFD: Fusion feature decoder.

Figure 2 Framework of stacked fruits detection algorithm under poor lighting based on RGB-D visual saliency



Note: Vec: Alignment feature vector; Evec: Enhancement vector; v_i: Edge enhancement vector; f_t: Common edge feature; f_c: Features after recalibration once; f_{rec}: Features after recalibration twice; w_i: Overall attention vector; Pr_i: Purified RGB feature.

Figure 3 Structure of depth-weighted preprocessing module

$$Fcal(d5') = F_{UP}^2\left(DConv_{3\times3}\left(F_{DN}^2(d5' + f_t)\right)\right) \quad (2)$$

where, Fcal represents a recalibration process. $DConv_{3\times3}$ means the dilated convolution kernel is 3×3, the step is 1, and the expansion rate is 2, including BatchNorm and ReLU activation. F_{UP}^2 and F_{DN}^2 indicate that the bilinear up-sampling and down-sampling operations are to 2 and 1/2 times the original size, respectively.

Two recalibrations were performed in this study for performance and efficiency tradeoffs, as shown in Equation (3).

$$f_c = Fcal(d5'), \quad f_{rec} = Fcal(f_c) \quad (3)$$

where, f_c and f_{rec} represent the features after recalibration once and twice, respectively. Finally, the overall attention vector w_i is obtained by combining f_{rec} and f_t, as shown in Equation (4).

$$w_i = BConv_{3\times3}(f_{rec} + f_t), \quad i = 1,2,\ldots,5 \quad (4)$$

where, $BConv_{3\times3}$ refers to 3×3 convolution with BatchNorm and Sigmoid activation.

Finally, the output of the depth-weighted preprocessing module was obtained, the purified RGB feature Pr_i, as Equation (5).

$$Pr_i = r_i + d_i * v_i * w_i, \quad i = 1,2,\ldots,5 \quad (5)$$

Unlike the DFM algorithm[15], which directly obtains the saliency maps through the depth weighting module, in this study, it was used as a preprocessing module to obtain purified RGB features. As shown in Figure 2, in the post-processing of the algorithm proposed, purified RGB features are used to replace the original input RGB features in order to improve the quality of the input image pairs.

2.2.2 Multi-scale progressive fusion

In order to effectively retain the modality-specific features in the process of multimodal feature fusion, increase the information interaction between branches of different scales, and fully cover the context information, so as to improve the robustness of the algorithm to light changes, a multi-scale progressive fusion (MPF) module is proposed. The specific structure of the MPF module is shown in Figure 4.

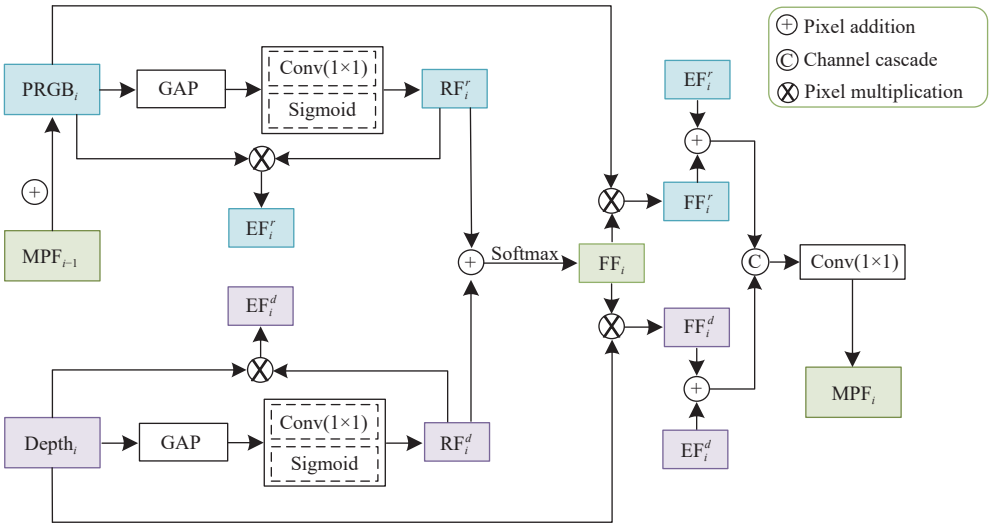
As shown in Figure 2, the input of the MPF module is the purified RGB features and depth features of each layer. Starting from the second branch, the output of the previous MPF module needs to be integrated into the purified RGB features before being input into the MPF module to achieve progressive fusion.

As shown in Figure 4, the MPF module first uses GAP for input features to obtain global statistics in RGB and depth images. Then, the two feature vectors into the 1×1 convolution layer and sigmoid activation function are used to obtain the reserved features RF_i^r and RF_i^d of the two modes, and then the reserved features are multiplied with the input features to obtain the modal enhancement features EF_i^r and EF_i^d, which represent the exclusive attributes of RGB features and depth features, respectively.

In addition, RF_i^r and RF_i^d aggregate through the Softmax function to retain the useful feature channels from RGB streams and depth streams, and then normalize them so that their output range is 0-1[17], thus obtaining the fused feature FF_i. The fusion features are multiplied with the input features of the two modes to obtain the

enhanced fusion features FF_i^r and FF_i^d. The output of the MPF module can be obtained by adding them with the enhanced features of the respective modes, and then cascading the results and entering 1×1 convolution, that is, the progressive fusion feature MPF_i.

The MPF module is characterized by the fact that from the first branch, the output of the previous MPF module will participate in the processing of the latter MPF module throughout the process, rather than a simple result cascade. The progressive fusion method integrates local and global features more effectively, increases the interaction between different branches, and can effectively retain valuable modal exclusive features while obtaining fusion features, preventing feature loss, maximizing the utilization of input features, and improving detection accuracy.



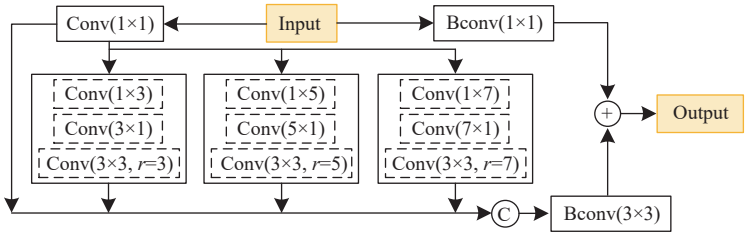
Note: RF_i^r: Reserved features of RGB mode; RF_i^d: Reserved features of Depth mode; EF_i^r and EF_i^d: Modal enhancement features; FFi: Fused feature; FF_i^r and FF_i^d: enhanced fusion features.

Figure 4 Structure of multi-scale progressive fusion module

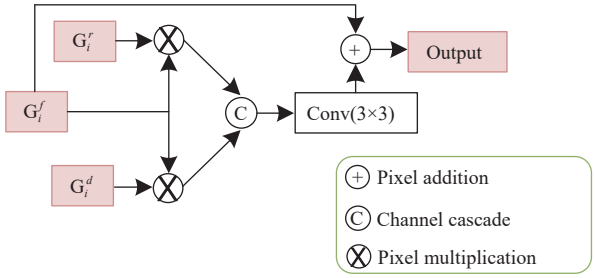
2.2.3 Combined decoders

As shown in Figure 2, the combined decoders used in this study are composed of Cascaded Partial Decoder (CPD)[18] and Fusion Feature Decoder (FFD)[19].

As shown in Figure 2, the input of the CPD module is the output of each layer of the purified RGB feature branch, depth feature branch, and fusion feature branch. The detailed structure of the CPD module is shown in Figure 5a. After the input features are dimensionally reduced by 1×1 convolution, it undergoes parallel spatially separable dilated convolution blocks with dilation rates of 3, 5, and 7, and the results are cascaded with the dimensionality reduction features and then input into Bconv3×3 convolutional layer, that is, 3×3 convolution with BatchNorm and ReLU activation. Finally, Bconv1×1 convolution is processed on the original input features, and the result is added to the Bconv3×3 convolution processing result and output. By using spatially separable dilated convolution blocks, the receptive fields can be improved without increasing the number of parameters. The multi-branch structure is used, each branch captures a receptive field, and finally the receptive field information is fused through cascading operations, which can achieve the effect of mimicking human vision, thereby improving the accuracy of the algorithm.



a. Structure of cascaded partial decoder



b. Structure of fusion feature decoder

Note: G_i^r: the outputs of the CPD module on the purified RGB feature branch; G_i^d: the outputs of the CPD module on the depth feature branch; G_i^f: the outputs of the CPD module on the fusion feature branch.

Figure 5 Combined decoders

Figure 2 shows that the inputs of the FFD module are the outputs of the CPD module on the purified RGB feature branch, the depth feature branch, and the fusion feature branch, which are recorded as G_i^r, G_i^d, and G_i^f, respectively. The detailed structure of the FFD module is shown in Figure 5b. First, G_i^r, G_i^d, with G_i^f respectively are multiplied, then the results are cascaded and 3×3 convolution is input. The convolution result and G_i^f are added to get the output of the FFD module. It can be seen from Figure 2 that the FFD module only appears in the fusion branch decoding stage. Its function is to integrate the modal exclusive features in the purified RGB feature branch and the depth feature branch into the fusion feature branch to obtain rich complementary multimodal information.

2.2.4 Multi-branch hybrid supervision

As shown in Figure 2, in the decoding stage, the purified RGB feature decoding branch, the depth feature decoding branch, and the fusion feature decoding branch will all generate an initial saliency map, which are recorded as SMr, SMd, and SMf, respectively. In this study, a mixed supervision method consisting of binary cross-entropy (BCE) loss and intersection-over-union (IOU) loss[20] was used for each branch, and the supervision was carried out by ground truth maps. The loss of each branch is shown in Equation (6).

$$\begin{cases} L_r = \frac{1}{2}\left(L_{\text{bce}}\left(SM\text{r}, GT\right) + L_{\text{iou}}\left(SM\text{r}, GT\right)\right) \\ L_d = \frac{1}{2}\left(L_{\text{bce}}\left(SM\text{d}, GT\right) + L_{\text{iou}}\left(SM\text{d}, GT\right)\right) \\ L_f = \frac{1}{2}\left(L_{\text{bce}}\left(SM\text{f}, GT\right) + L_{\text{iou}}\left(SM\text{f}, GT\right)\right) \end{cases} \tag{6}$$

where, L_r represents the loss of the purified RGB feature decoding branch, L_d represents the loss of the depth feature decoding branch, L_f represents the loss of the fusion feature decoding branch, L_{bce} represents the loss of binary cross-entropy, and L_{iou} represents the loss of joint intersection.

The total loss of the proposed algorithm is as Equation (7).

$$L = L_r + L_d + L_f \tag{7}$$

The hybrid supervision method can make pixel-level BCE loss and mapping-level IOU loss complement each other, accelerate algorithm convergence, and reduce the impact of adverse factors such as reflection and low illumination, so as to improve the accuracy of saliency prediction.

3 Experiments and discussion

3.1 Experiment platform and parameter setting

The experimental running environment is a 64-bit Ubuntu 20.04 operating system, 64 GB memory, and one Geforce GTX 3090 GPU. The deep learning framework on which this algorithm is based is Pytorch[21], and the input RGB and depth image pairs are adjusted to 352×352 resolution. Using the Adam optimization model[22], the initial learning rate is set to 1e-4 and decreases by 10 times every 60 epochs. The batch size is set to 10, and the model has trained 100 epochs.

3.2 Evaluation indices

This paper uses four evaluation indicators to evaluate the performance of the models, including the precision-recall (*P*-*R*) curve, *F*-value[23], mean absolute error (MAE)[24], and *S*-value[25].

The *P*-*R* curve is used to calculate the accuracy and recall of the feature maps. When the output images are binarized, the threshold values are selected from 0 to 255, and each time the threshold value is acquired, a set of corresponding precision values and recall values can be calculated for all output images. Finally, the precision values and recall values of all images under the threshold are averaged to obtain 256 pairs of *P* values and *R* values. The recall value is abscissa and the precision value is ordinate. A curve graph is drawn to obtain the *P*-*R* curve. The closer the *P*-*R* curve is to the upper right corner, the better the algorithm performance.

F-value[23] is the weighted harmonic average of recall rate and accuracy rate under non-negative weight, and its calculation method is shown in Equation (8).

$$F = \frac{(1+\beta^2)PR}{\beta^2 P + R} \tag{8}$$

where, β is a non-negative weight to measure recall and accuracy. According to the experience of many salient target detection tasks, it is usually set as β^2=0.3, increasing the weight value of accuracy. In this study, the maximum *F*-value is used to represent the best performance of the algorithm.

MAE[24] represents the mean value of the absolute error between the predicted value and the true value. The range is that when the predicted value is completely consistent with the true value, it is equal to 0, which is a perfect model. The greater the error between the predicted value and the true value, the greater the MAE value. The MAE calculation method is shown in Equation (9).

$$\text{MAE} = \frac{1}{m}\sum_{i=1}^{m}|y_i - f(x_i)| \tag{9}$$

where, m represents the number of samples, $f(x)$ represents the predictive value of the model, and y represents the true value. The smaller the MAE value, the higher the similarity between the predicted value and the true value, the less the background noise, and the better the overall performance of the algorithm.

S-value[25] focuses on evaluating the structural information of saliency maps, which is closer to the human visual system than *F*-value. It mainly calculates the structural similarity of object perception and region perception between the predicted value and the true value. The calculation method of *S*-value is shown in Equation (10).

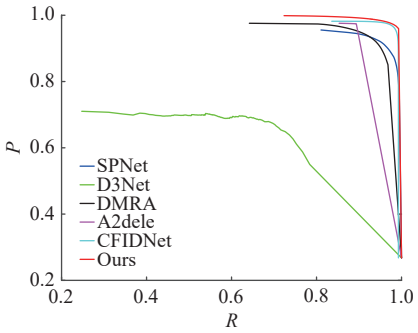
$$S_\lambda = \lambda \times S_o + (1-\lambda) \times S_a \tag{10}$$

where, $\lambda \in [0,1]$ is a balance parameter, usually taken as 0.5. S_o and S_a represent the structural similarity of target perception and area perception, respectively. The larger the *S*-value, the smaller the structural error of the saliency map and ground truth map, and the better the performance of the algorithm.

3.3 Quantitative evaluation of experiment results

Figure 6 shows the *P*-*R* curves of state-of-the-art RGB-D SOD methods on the self-made stacked fruit test dataset in recent years. The algorithm proposed in this study is closest to the upper right corner of the coordinate system. Especially when the recall rate is in the range of 0.8-1.0, the accuracy value is larger than other algorithms. Therefore, the algorithm in this paper is superior to other algorithms on the whole and has high detection reliability.

Table 2 lists the other three objective evaluation criteria. The best value is shown in bold. It can be seen that the *S*-value, *F*-value, and MAE value of the algorithm proposed in this study have absolute advantages over other algorithms on stacked fruit datasets. From the comprehensive objective evaluation indicators, the spatial structure of the saliency maps predicted by the algorithm proposed in this study is closer to that of the ground truth maps.



Note: P is Precision, R is Recall, and Ours means the method proposed in this study. Same below.

Figure 6 P-R curves of comparison algorithms on stacked fruit dataset

3.4 Qualitative evaluation of experiment results

The visual comparison results between the algorithm proposed in this study and other comparison algorithms on the stacked fruit dataset are shown in Figures 7-9, where Figure 7 shows the detection results in the case of overexposure and reflection, Figure 8 shows the detection results in the case of non-uniform illumination, and Figure 9 shows the detection results in the case of low illumination. It can be seen from the detection results of different algorithms on various images in Figures 7-9 that under poor lighting conditions such as overexposure, non-uniform illumination, and low illumination, the algorithm proposed in this study is more robust, and can extract clear foreground target edges, resulting in more complete and smooth significant areas, good retention of target fruit details, and high similarity with the ground truth maps. In summary, the algorithm proposed in this study has strong anti-interference ability, is less affected by lighting conditions, and has high positioning and detection accuracy for various types of fruit targets. Therefore, the algorithm proposed in this study has good detection performance for the same kind of stacked fruits.

Table 2 Evaluation criteria of comparison algorithms on stacked fruit dataset

Evaluation criteria	Algorithms					
	SPNet[19]	D3Net[26]	DMRA[27]	A2dele[28]	CFIDNet[29]	Ours
$S\uparrow$	0.915	0.674	0.915	0.936	0.976	**0.979**
$F\uparrow$	0.934	0.681	0.939	0.974	0.989	**0.992**
MAE↓	0.042	0.167	0.043	0.029	0.011	**0.006**

Note: Ours means the method proposed in this study. Same below.

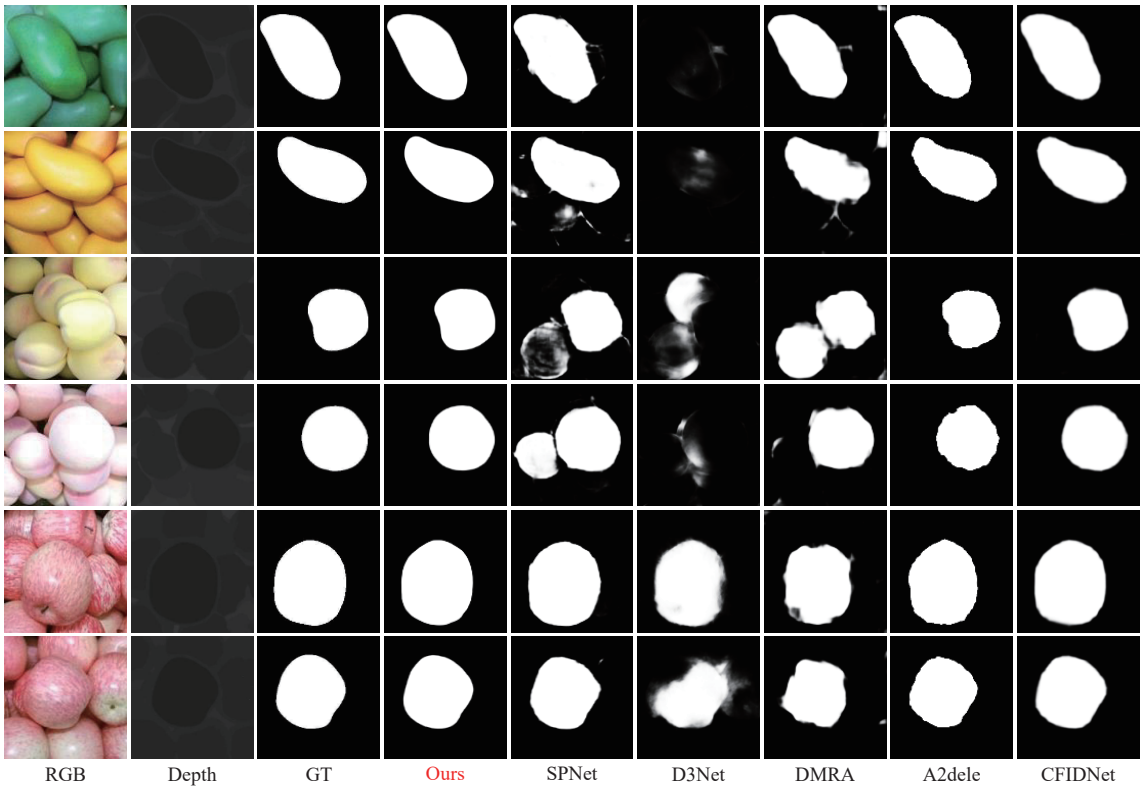


Figure 7 Comparison of detection results of stacked fruits with overexposure and reflection

3.5 Module ablation experiments

In this study, the DES public dataset[30] with poor lighting conditions is taken as an example to verify the impact of the three modules on the algorithm performance.

The specific operation is to remove a module separately and keep other settings unchanged for training. The baseline represents the basic network without all modules, while "w/o DWP" means to remove the depth-weighted preprocessing module, "w/o MPF" means to remove the multi-scale progressive fusion module, and "w/o combined decoders" means to remove the decoding part. The P-R curves of the DES dataset are shown in Figure 10. It can be seen from Figure 10 that removing any of the three modules will affect the algorithm's performance. Table 3 lists the comparison between the objective evaluation criteria obtained on the DES dataset after removing any module and the evaluation criteria of the complete algorithm proposed in this study. It can be seen that the removal of any module will lead to a decrease in F-value and S-value, and an increase in MAE value. Therefore, it can be proved that the three modules included in this algorithm help improve the detection accuracy.

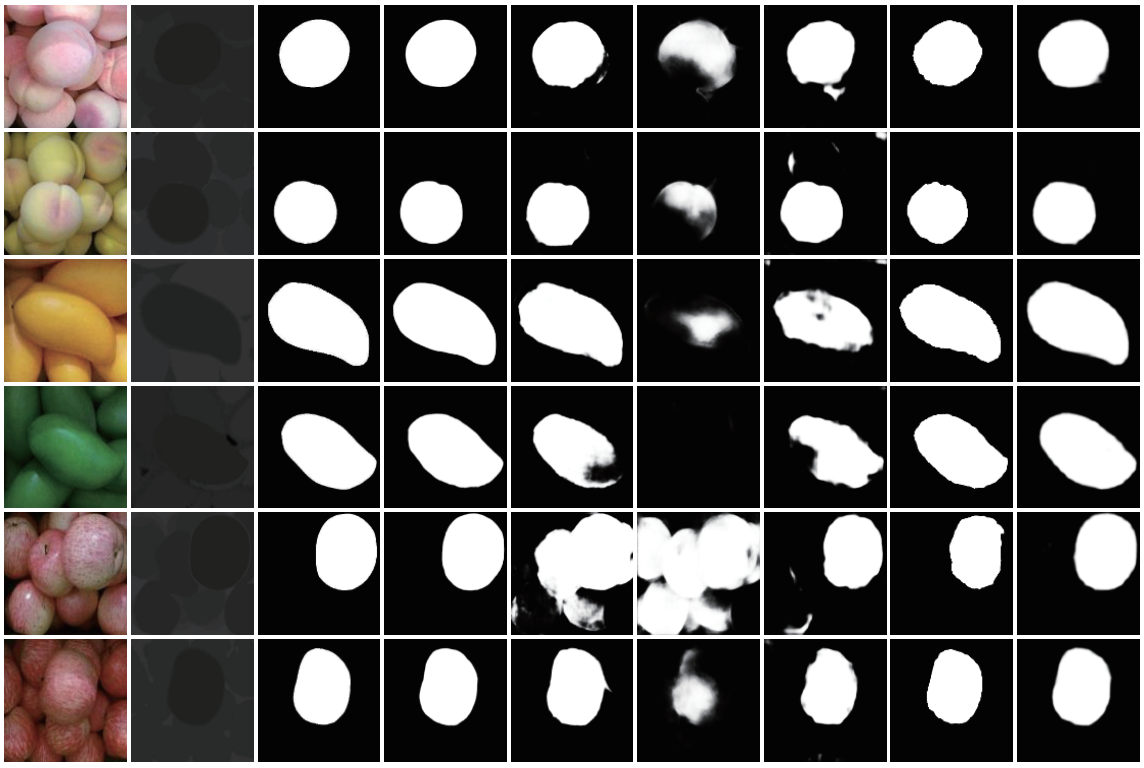


Figure 8 Comparison of detection results of stacked fruits with non-uniform illumination

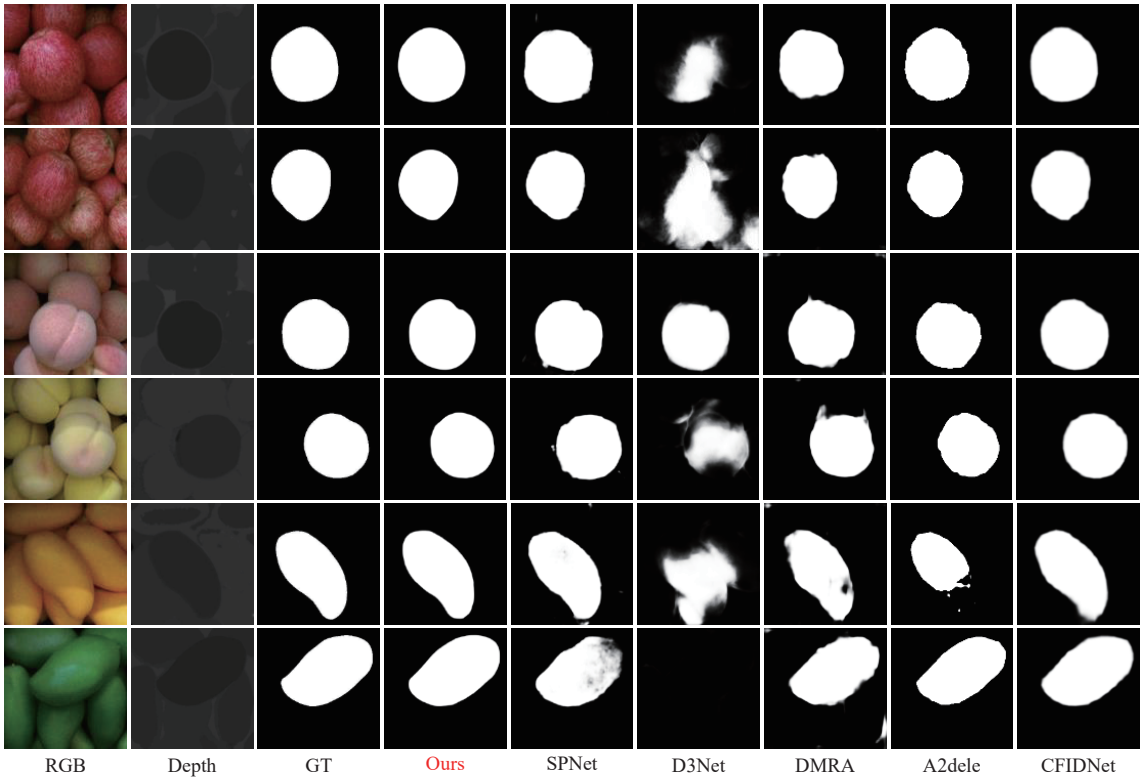


Figure 9 Comparison of detection results of stacked fruits with low illumination

4 Conclusions

The method for detecting stacked fruits under poor illumination based on RGB-D visual saliency was proposed, which realizes the detection of salient objects of the same kind of stacked fruits, can help the robot quickly lock the optimal grasping target, avoids losses caused by fruit tumbling and falling, and is conducive to promoting the intelligent process of fruit production lines. Through the depth-weighted preprocessing module, the original input RGB features are purified, which reduces the influence of low-quality depth maps on the detection effect. The multi-scale progressive fusion module is used to effectively retain the exclusive features of modes in the process of multi-modal feature fusion, increase the information interaction between branches of different scales, fully cover the context information, and maximize the utilization of input features, thereby improving the robustness of the model to poor lighting conditions. The initial saliency maps generated by the multi-

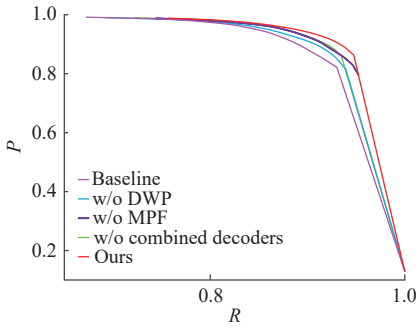


Figure 10 *P-R* curves of module ablation experiments on the DES dataset

Table 3 Evaluation criteria of module ablation experiments on the DES dataset

Evaluation criteria	Module ablation experiments				
	Baseline	w/o DWP	w/o MPF	w/o combined decoders	Ours
S↑	0.906	0.920	0.915	0.921	**0.926**
F↑	0.929	0.936	0.932	0.937	**0.939**
MAE↓	0.025	0.922	0.024	0.023	**0.020**

branch combined decoders adopt a hybrid supervision method to balance the differences of different modal features and finally obtain saliency maps with clear edges of complete content. On the self-made stacked fruit dataset, the proposed method is qualitatively and quantitatively compared with five state-of-the-art RGB-D SOD methods, and the experimental results show that the detection results of the proposed algorithm are the closest to the ground truth maps and have certain advantages in four objective evaluation indicators. The detection frame rate of the algorithm proposed in this study is 8.5 fps (the detection image resolution is 352×352). Subsequent work will focus on improving the continuous detection speed of the algorithm, so as to better meet the real-time requirements of fruit sorting tasks.

[References]

[1] Dewi T, Risma P, Oktarina Y. Fruit sorting robot based on color and size for an agricultural product packaging system. Bulletin of Electrical Engineering and Informatics, 2020; 9(4): 1438–1445.

[2] Fan S X, Liang X T, Huang W Q, Zhang V J L, Pang Q, He X, et al. Real-time defects detection for apple sorting using NIR cameras with pruning-based YOLOV4 network. Computers and Electronics in Agriculture, 2022; 193: 106715.

[3] Kumaravel G, Ilankumaran V, Al Maqrashi S A A, Al Yaaqubi M K S. Automated date fruits sorting machine using fuzzy logic controller. International Journal of Recent Technology and Engineering, 2019; 8(4): 1089–1093.

[4] Wang Y, Wang Z T. Salient object detection based on deep network. Electronic Measurement Technology, 2019; 42(21): 101–104. (in Chinese)

[5] Tan D N, Liu Y, Yao L B, Ding Z R, Lu X Q. Semantic segmentation of multi-source remote sensing images based on visual attention mechanism. Journal of Signal Processing, 2022; 38(6): 1180–1191. (in Chinese)

[6] Lyu P F, Wang Y, Wang S Q, Yu X S, Wu C D. Optic disc detection based on visual saliency in fundus image. Journal of Image and Graphics, 2021; 26(9): 2293–2304. (in Chinese)

[7] Arivazhagan S, Shebiah R N, Harini R, Swetha S. Human action recognition from RGB-D data using complete local binary pattern. Cognitive Systems Research, 2019; 58: 94–104.

[8] Liu D, Hu Y S, Zhang K, Chen Z Z. Two-stream refinement network for RGB-D saliency detection. In: 2019 IEEE International Conference on Image Processing (ICIP), Taipei: IEEE, 2019; pp.3925–3929.

[9] Zhao J X, Cao Y, Fan D P, Cheng M M, Li X Y, Zhang L. Contrast prior and fluid pyramid integration for RGBD salient object detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition, Long Beach, 2020; pp.3922–3931. doi: 10.1109/CVPR.2019.00405.

[10] Singh S K, Srivastava R. SL-Net: Self-learning and mutual attention-based distinguished window for RGBD complex salient object detection. Neural Computing and Applications, 2022; 35: 595–609.

[11] Das D K, Shit S, Ray D N. Depth-guided two-way saliency network for 2D images. In: Advanced computational paradigms and hybrid intelligent computing. Springer, 2022; pp.61–71. doi: 10.1007/978-981-16-4369-9_7.

[12] Ju R, Ge L, Geng W J, Ren T W, Wu G S. Depth saliency based on anisotropic center-surround difference. In: 2014 IEEE International Conference on Image Processing (ICIP), Paris: IEEE, 2014; pp.1115–1119. doi: 10.1109/ICIP.2014.7025222.

[13] Peng H, Li B, Xiong W H, Hu W M, Ji R R. RGBD salient object detection: A benchmark and algorithms. In: Proceedings of the European Conference on Computer Vision (ECCV), 2014; pp.92–109. doi: 10.1007/978-3-319-10578-9_7.

[14] Gao S H, Cheng M M, Zhao K, Zhang X Y, Yang M H, Torr P. Res2Net: A new multi-scale backbone architecture. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2019; 43(2): 652–662.

[15] Zhang W B, Ji G P, Wang Z, Fu K R, Zhao Q J. Depth quality-inspired feature manipulation for efficient RGB-D salient object detection. In: Proceedings of the 29th ACM International Conference on Multimedia, 2021; pp.731–740. doi: 10.1145/3474085.345240.

[16] Zhang P P, Wang D, Lu H C, Wang H Y, Ruan X. Amulet: Aggregating multi-level convolutional features for salient object detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), Venice: IEEE, 2017; pp.202–211. doi: 10.1109/ICCV.2017.31.

[17] Ji W, Li J J, Yu S, Zhang M, Piao Y R, Yao S Y. Calibrated RGB-D salient object detection. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville: IEEE, 2021; pp.9466–9476. doi: 10.1109/CVPR46437.2021.00935.

[18] Wu Z, Su L, Huang Q M. Cascaded partial decoder for fast and accurate salient object detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach: IEEE, 2019; pp.3902–3911. doi: 10.1109/CVPR.2019.00403.

[19] Zhou T, Fu H Z, Chen G, Zhou Y, Fan D P, Shao L. Specificity-preserving RGB-D saliency detection. In: 2021 IEEE/CVF International Conference on Computer Vision (ICCV), Montreal: IEEE, 2021; pp.4661–4671. doi: 10.1109/ICCV48922.2021.00464.

[20] Qin X B, Zhang Z C, Huang C Y, Gao C, Dehghan M, Jagersand M. BASNet: Boundary-aware salient object detection. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach: IEEE, 2019; pp.7471–7481. doi: 10.1109/CVPR.2019.00766.

[21] Steiner B, Devito Z, et al. PyTorch: An imperative style, high performance deep learning library. arXiv, 2019; arXiv: 1912.01703. In Press. doi: 10.48550/arXiv.1912.01703.

[22] Kingma D, Ba J. Adam: A method for stochastic optimization. arXiv, 2014; arXiv: 1412.6980. doi: 10.48550/arXiv.1412.6980.

[23] Achanta R, Hemami S, Estrada F, Susstrunk S. Frequency-tuned salient region detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition, Miami: IEEE, 2009; pp.1597–1604. doi: 10.1109/CVPR.2009.5206596.

[24] Borji A, Cheng M M, Jiang H Z, Li J. Salient object detection: A benchmark. IEEE Transactions on Image Processing, 2015; 24(12): 5706–5722.

[25] Fan D P, Cheng M M, Liu Y, Li T, Borji A. Structure-measure: A new way to evaluate foreground maps. In: 2017 IEEE International Conference on Computer Vision (ICCV), Venice: IEEE, 2017; pp.4558–4567. doi: 10.1109/ICCV.2017.487.

[26] Fan D P, Lin Z, Zhang Z, Zhu M L, Cheng M M. Rethinking RGB-D salient object detection: Models, data sets, and large-scale benchmarks. IEEE Transactions on Neural Networks and Learning Systems, 2020; 32(5): 2075–2089.

[27] Piao Y J, Ji W, Li J J, Zhang M, Lu H C. Depth-induced multi-scale recurrent attention network for saliency detection. In: 2019 IEEE/CVF International Conference on Computer Vision (ICCV), Seoul: IEEE, 2020; pp.7253–7262. doi: 10.1109/ICCV.2019.00735.

[28] Piao Y R, Rong Z K, Zhang M, Ren W S, Lu H C. A2dele: Adaptive and attentive depth distiller for efficient RGB-D salient object detection. In: IEEE/CVF Conference on Computer Vision and Pattern Recognition, Seattle, 2020; pp.9057–9066. doi: 10.1109/CVPR42600.2020.00908.

[29] Chen T Y, Hu X G, Xiao J, Zhang G F, Wang S J. CFIDNet: Cascaded feature interaction decoder for RGB-D salient object detection. Neural Computing and Applications, 2022; 34(10): 7547–7563.

[30] Cheng Y P, Fu H Z, Wei X X, Xiao J J, Cao X C. Depth enhanced saliency detection method. In: Proceedings of International Conference on Internet Multimedia Computing and Service, 2014; pp.23–27. doi: 10.1145/2632856.2632866.