Salient Object Detection Based on Improved Pyramid Pooling Network

MengHuai Xiao
School of Computer and Information Engineering
Fuyang Normal University
Anhui, China
sduxmh@163.com

Yue Wang
School of Computer and Information Engineering
Fuyang Normal University
Anhui, China
572363120@qq.com

Ya Wang*
School of Computer and Information Engineering
Fuyang Normal University
Anhui, China
fync_wy80@163.com
*corresponding author

Jun Huang
School of Computer and Information Engineering
Fuyang Normal University
Anhui, China
jjun1110@163.com

2023 16th International Conference on Advanced Computer Theory and Engineering (ICACTE) | 979-8-3503-1753-4/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICACTE59887.2023.10335231

Abstract—**Salient object detection is an important task in computer vision, which is to segment visually prominent objects from images accurately. However, image segmentation for small objects and complex background objects is still a challenging task. In order to improve the segmentation effect, a new pyramid-pool network method based on PoolNet network is proposed in this paper. The network, called CDSPNet, integrates Convolutional Block Attention Module (CBAM) and Deep Supervision (DS), wherein the convolutional block attention module can effectively improve the representation ability of feature maps. Depth supervision adds side output to better blend rich deep features. The experimental results show that CDSPNet achieves better salient object detection performance by comparing with other 8 models in F-measure and MAE on 6 public datasets.**

Keywords—Salient Object Detection, CDSPNet, Convolutional Block Attention Module, Deep Supervision

I. Introduction

Distinct from other similar segmentation tasks such as semantic segmentation, salient object detection aims to segment the most visually attractive objects in an image. It is widely used in image segmentation, photo synthesis, behavior recognition and visual tracking[1]. Traditional methods[2] mostly rely on hand-made features to capture local details and global context information of images, but lack of high-level semantic information limits their ability to detect overall significant targets in complex scenes. However, with the development of deep convolutional neural networks (CNNs), especially the emergence of full convolutional networks (FCN) in image segmentation, the development of multi-scale spatial details has been greatly promoted, and salient object detection has been significantly improved.

Liu et al.[3] proposed PoolNet network. Based on the feature pyramid network (FPNs), the network added two main modules: global guidance module (GGM) and feature aggregation module (FAM), which were used to integrate global significance features to improve the detection performance of targets. However, when detecting small targets and complex foreground objects that overlap with the image boundary, the PoolNet network can not integrate image feature information well, which leads to the degradation of salient object detection performance. To solve this problem, we propose a new network CDSPNet that improves PoolNet. The main contributions are as follows:

We summarize the contributions of our work as follows:

The Convolutional Block Attention Module[4] (CBAM) is introduced into the network to better locate the most significant area of the image.

Depth Supervision[5] (DS) is used to conduct feature fusion from deep layer to shallow layer by means of multi-level features, so that the shallower side output layer can fuse more rich deep features, and thus improve the network object detection performance.

II. Network Structure

By integrating the convolutional block attention module and deep supervision, CDSPNet can better extract the feature information of the image and retain more details of the edge information to improve the effect of network segmentation.

A. Overall Structure of CDSPNet

The main structure of CDSPNet is based on FPNs structure, which is designed in top-down and bottom-up ways. FPNs is widely used in many visual tasks, including salient object detection, because of its strong ability to combine multi-level features. The network uses ResNet-50[6] as our backbone, and then introduces the global guidance module (GGM), which is established in the bottom-up path of the network. The high-level semantic information extracted by GGM and semantic information extracted at different levels of the network are fused in the feature aggregation module (FAM). In order to obtain a more accurate salient maps, we introduce CBAM attention mechanism to better locate the location information of the image, and use DS to retain the semantic information of each stage, and utilize multi-level feature fusion to improve the object detection performance of the network. The overall architecture of CDSPNet network model is shown in Fig. 1.

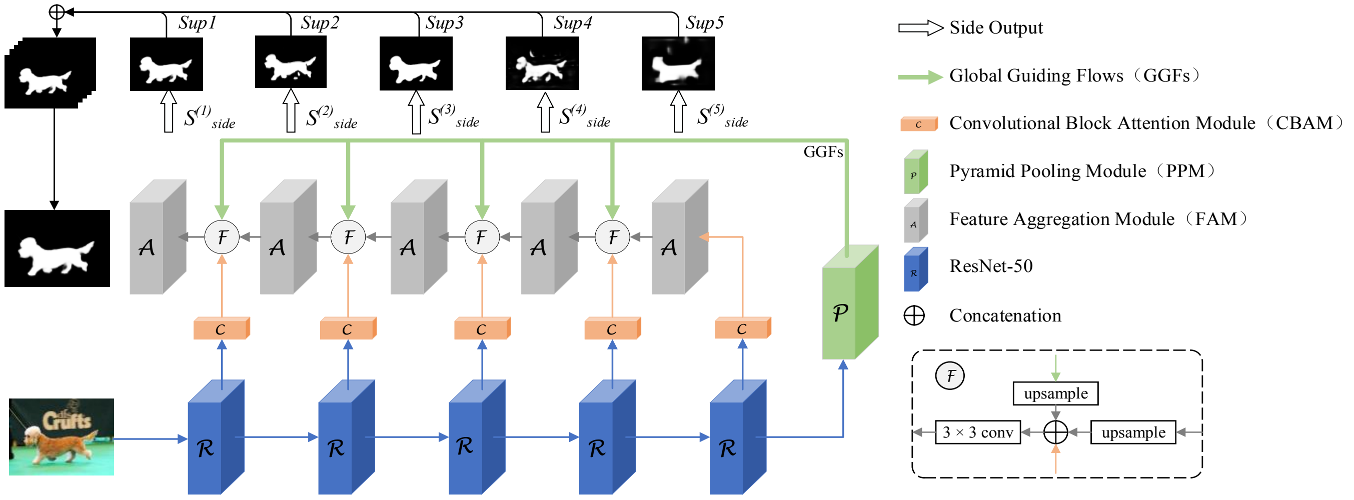

Fig. 1. Overall architecture of CDSPNet network model

B. Convolution Block Attention Module

CBAM is a technique used to enhance network models' attention to information in feature maps channels and spatial dimensions. It consists of two sub-modules: Channel Attention Module (CAM) and Spatial Attention Module (SAM), which mainly introduce attention mechanism in channel and spatial dimension respectively. CAM is mainly used to adjust the feature weights of different channels, while SAM is mainly used to adjust the feature weights of different spatial positions, so as to guide the model to pay better attention to the important features of the image and further improve the detection performance of the network model, as shown in Fig. 2.

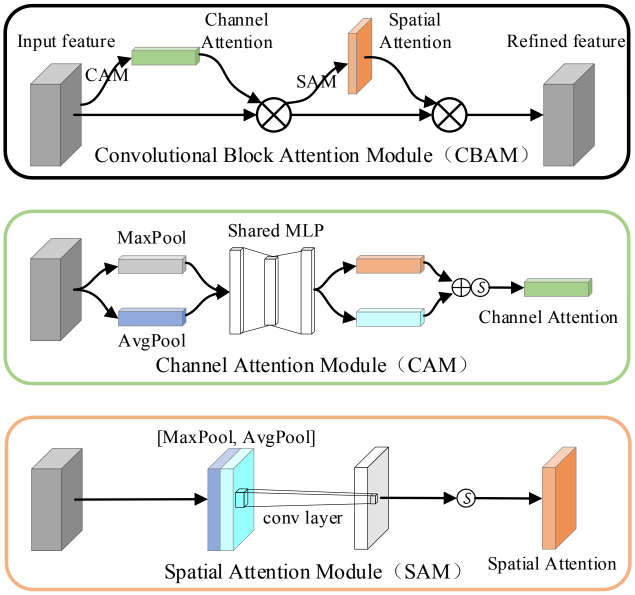

Fig. 2. Convolution Block Attention Module

C. Deep Supervision

DS is designed to improve the performance and robustness of network models by adding supervisory signals at different network levels. DS is similar to HED[7], but different from traditional methods, this method introduces multiple side output branches into the network, each branch is used to detect edge information of different scales, and fuse feature information of different scales. In the training process, using multiple loss functions to guide the learning of different hierarchical networks can effectively improve the performance of the network model and avoid problems such as the disappearance of gradient. We visualized five side output branches in the network (s-out1, s-out2, s-out3, s-out4, s-out5) and compared them with the corresponding ground truth (GT), as shown in Fig. 3.

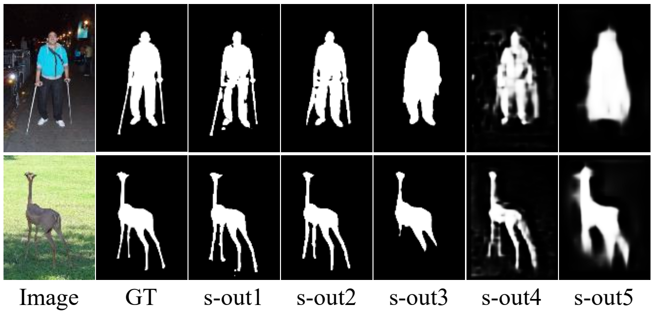

Fig. 3. 5 branch side output compared with GT

As shown in Fig. 3, in the bottom-up path of the network, the saliency maps obtained by the bottom network pays more attention to the overall location information of the target, while the top network, when receiving the feature maps from the bottom network, can effectively supplement the deficiency of the top network in scale and details by integrating the initial features obtained by ResNet-50.

III. EXPERIMENTAL RESULTS AND ANALYSIS

The CDSPNet backbone network uses ResNet-50, which is initialized by loading ResNet-50's pre-training model on the ImageNet dataset. The experiment was trained in PyTorch 1.13 environment, the network trained 24 epoches in total. Adam optimizer was used. The weight attenuation was set to 5e-4, the initial learning rate was set to 5e-5, and the learning rate was multiplied by 0.1 after the 15th epoch. The dataset was enhanced by a simple random horizontal flip. During training and testing, the size of the input image remains the same. First, an ablation experiment was conducted to demonstrate the effect of the proposed method on the performance of the entire network model. Finally, according to the two evaluation indexes, the model was compared with UCF[8], Amulet[9], PiCANet[10], PiCANet-R[10], CPD[11], PoolNet[3], BASNet[12] and U^{2}Net[13].

A. Datasets

Training dataset: Our CDSPNet network uses DUTS-TR in the DUTS dataset as the training set for training, which contains a total of 10553 images. At present, DUTS-TR is the most widely used and the largest dataset for saliency object detection training.

Evaluation datasets: We used six commonly used publicly available datasets to evaluate CDSPNet networks, including: DUT-OMRON, DUTS-TE, HKU-IS, ECSSD, PASCAL-S, and SOD datasets.

Loss function Loss: During training, depth supervision similar to HED is used. The training loss formula is defined as formula (1) :

$$\mathcal{L} = \sum_{m=1}^{M} w_{side}^{(m)} \ell_{side}^{(m)} + w_{fuse}\, \ell_{fuse} \qquad (1)$$

In formula (1), $\ell_{side}^{(m)}$ (M=5, indicating that Sup1-Sup5 in Fig. 1) is the loss of side output $S_{side}^{(m)}$, and ℓ_{fuse} is the loss of the final fusion output saliency maps. $w_{side}^{(m)}$ and w_{fuse} are the weight of each saliency maps loss. For each saliency maps, the loss is calculated using the standard binary cross entropy method, as shown in formula (2) :

$$\ell = -\sum_{(r,c)}^{(H,W)} \left[\begin{array}{c} P_{G(r,c)} log P_{S(r,c)} + \\ \left(1 - P_{G(r,c)}\right) \log\left(1 - P_{S(r,c)}\right) \end{array} \right] \qquad (2)$$

In formula (2), (r, c) is pixel coordinate, and (H, W) is image size: height and width. $P_{G(r,c)}$ and $P_{S(r,c)}$ represent the pixel values of ground truth (GT) and predicted saliency maps, respectively. In the process of network test, we choose the final fusion result as the final saliency maps.

B. Evaluation Index

Salient object detection is an important research direction in the field of computer vision, which poses certain challenges to the accuracy and stability of evaluation algorithms. Therefore, the design of effective evaluation index is very important to measure the performance of the algorithm and evaluate the effect of the algorithm. The following are two commonly used salient object detection evaluation indexes, *maxF$_\beta$* and *MAE*.

F-measure: F_β integrated Precision and Recall index, which is used to evaluate the comprehensive performance of the salient object detection algorithm. Its calculation formula is shown in equation (3) :

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \qquad (3)$$

We set β^2 in Equation (3) to 0.3 to evaluate the dataset, and obtain the maximum F_β (*maxF$_\beta$*) evaluation index of each dataset.

MAE: mean absolute error, defined as the mean of the absolute value difference between the actual value and the predicted value. It is a commonly used evaluation indicator to measure the absolute error between the predicted value and the actual value, and to measure the difference between the saliency object detection output saliency maps and the real saliency maps, regardless of the sign of the actual value. Its advantage is that it is more robust to outliers because it is not affected by noise. Its formula is defined as formula (4) :

$$MAE = \frac{1}{H \times W} \sum_{r=1}^{H} \sum_{c=1}^{W} |P(r,c) - G(r,c)| \qquad (4)$$

In formula (4), P and G are the probability maps of the salient object detection and the corresponding GT respectively, (H, W) and (r, c) are the (height, width) and the pixel coordinates.

C. Ablation Experiment

Ablation experiments were performed on the benchmark model, PoolNet, trained using the DUTS-TR dataset (10,553 images) and different combinations of CBAM and DS. Our model CDSPNet combines CBAM and DS methods. The experiment was tested on DUT-OMRON and SOD datasets. CBAM and DS were used for ablation experiments, and other parameters were consistent with PoolNet. The results of the ablation experiment are shown in Table 1.

TABLE I. ABLATION EXPERIMENT

Settings	**DUT-OMROM**		**SOD**	
	maxF$_\beta$	*MAE*	*maxF$_\beta$*	*MAE*
Baseline	0.804	0.055	0.865	0.109
CBAM	0.806	0.055	0.861	0.109
DS	0.807	0.056	0.869	**0.103**
CBAM+DS	**0.807**	**0.053**	**0.878**	0.106

According to the analysis of Table 1, only adding CBAM module can achieve slightly better results in the *maxF$_\beta$* evaluation index of DUT-OMRON dataset. Only DS method was used to obtain good performance in *maxF$_\beta$* evaluation index of two datasets, and the best performance was obtained in *MAE* evaluation index of SOD dataset. Finally, by introducing CBAM and DS into the benchmark model at the same time, both *maxF$_\beta$* and *MAE* evaluation indexes were further improved to achieve the best performance.

D. Experimental Result

Our model (Ours, i.e. CDSPNet) is qualitatively compared with 8 significant object detection methods[11]. In order to compare the experimental results fairly, the data of the other models were evaluated using the same evaluation code as the salient object detection maps provided by the authors, except that PoolNet was our benchmark model.

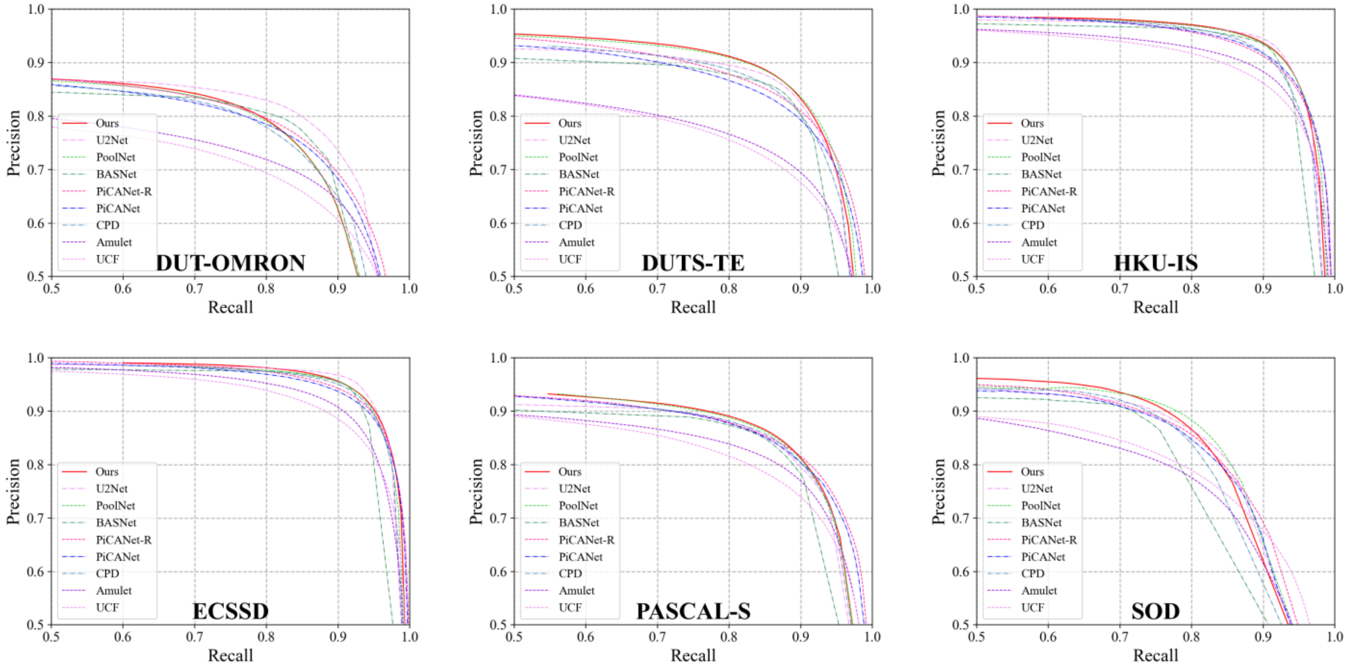

Fig. 4. P-R curve

TABLE II. THE EVALUATION INDEXES OF EACH ALGORITHM MODEL ARE COMPARED ON 6 COMMONLY USED DATASETS

Model	Training Datasets	DUT-OMRON		DUTS-TE		HKU-IS		ECSSD		PASCAL-S		SOD	
		$maxF_\beta$	*MAE*	$maxF_\beta$	*MAE*	$maxF_\beta$	*MAE*	$maxF_\beta$	*MAE*	$maxF_\beta$	*MAE*	$maxF_\beta$	*MAE*
UCF	MK	0.730	0.120	0.773	0.112	0.888	0.062	0.903	0.069	0.814	0.115	0.808	0.148
Amulet	MK	0.743	0.098	0.778	0.084	0.897	0.051	0.915	0.059	0.830	0.100	0.798	0.144
PiCANet	DUTS-TR	0.794	0.068	0.851	0.054	0.921	0.042	0.931	0.046	0.856	0.078	0.854	0.103
PiCANet-R	DUTS-TR	0.803	0.065	0.860	0.050	0.918	0.043	0.936	0.046	0.860	0.076	0.855	**0.103**
CPD	DUTS-TR	0.797	0.056	0.865	0.043	0.925	0.034	0.939	0.037	0.861	**0.071**	0.860	0.110
PoolNet	DUTS-TR	0.804	0.055	*0.882*	*0.039*	0.932	0.032	0.943	0.038	*0.866*	0.074	*0.865*	0.109
BASNet	DUTS-TR	0.805	0.056	0.860	0.047	0.928	0.032	0.942	0.037	0.856	0.076	0.851	0.112
U2Net	DUTS-TR	**0.823**	*0.055*	0.873	0.044	**0.935**	**0.031**	**0.951**	**0.033**	0.861	0.074	0.861	0.106
Ours	DUTS-TR	*0.807*	**0.053**	**0.883**	**0.038**	*0.933*	*0.032*	*0.944*	*0.037*	**0.868**	*0.072*	**0.878**	*0.106*

Note: Red bold is the best, blue italic is the second best, the larger the $maxF_\beta$, the smaller the *MAE*

Fig. 4 shows the P-R curve of our model (Ours red curve) on 6 datasets compared with 8 models, Ours is superior to other models on DUTS-TE, HKU-IS and SOD datasets.

Two commonly used evaluation indexes, $maxF_\beta$ and *MAE*, are used in 6 datasets to make qualitative comparison between the data evaluated by other models and Ours, and the results are shown in Table 2.

As can be seen from Table 2, Ours model achieved the best performance of 88.3%, 86.8% and 87.8% of $maxF_\beta$ evaluation indexes on DUTS-TE, PASCAL-S and SOD datasets, respectively. A minimum *MAE* of 0.053 and 0.038 was obtained on the DUT-OMRON and DUTS-TE datasets, respectively. And two evaluation indicators on other datasets achieved second-best performance, mainly slightly lower than the U^{2}Net network.

Below, the segmentation effects of the above 8 models are compared, as shown in Fig. 5.

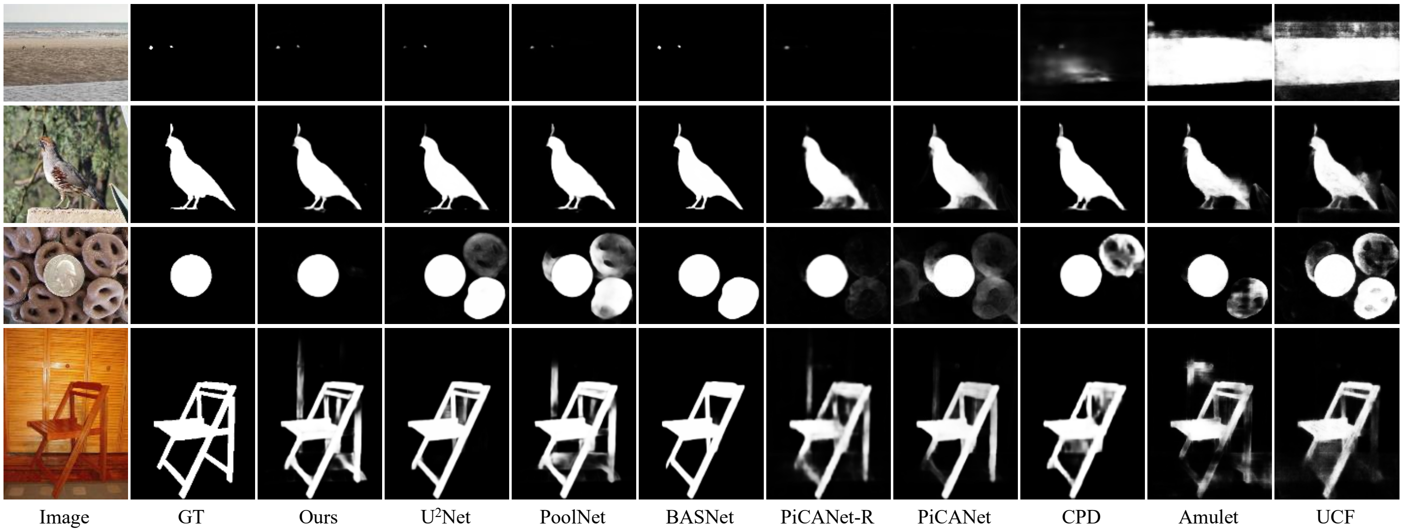

Fig. 5. Salient object detection results with 8 models

The first and second lines in Fig. 5 are the segmentation of small target and large target. Ours can accurately locate the target and carry out the segmentation. The third line shows the segmentation target in the case of complex background interference, other models are interfered with the background and segmentation fails, but our model gets perfect segmentation effect. In line 4, the segmentation target has a similar color to the surrounding environment, Ours can also segment the outline of the whole target relatively completely. Obviously, compared with other models, Ours can not only locate the overall details of small and large targets, but also can not be disturbed by complex background. This results in a saliency maps very close to ground truth.

IV. Conclusions

Based on the PoolNet network, we designed the CDSPNet network model and introduced CBAM and DS to enhance the segmentation effect of the network on the saliency maps. Experiments show that compared with other models, our model has better detection performance in 6 commonly used datasets. The overall structure of CDSPNet model uses the FPNs structure with ResNet-50 as the backbone. The FPNs structure can extract features from images of different scales and generate multi-scale feature representations, but there are also limitations. Using FPNs structure leads to long reasoning time and large memory consumption of our model. CDSPNet model introduces GGM module into FPNs structure to capture the global context information better, but neglects the problem that the differences and details between different targets cannot be accurately captured due to the characteristics of global guidance. To solve these problems of CDSPNet model, we will further improve the model from the aspects of network structure design and feature extraction in the future.

Acknowledgment

This research was funded by Anhui Provicial Overseas Study Visit Project (gxgwfx200050), Fuyang Normal University of Natural Science Research Project (2018kyqd0028), Fuyang Normal University of Science Research and Innovation Team (kytd202004), Anhui Graduate Graduate Innovation and Entrepreneurship Practice Project （2022cxcysj189).

References

[1] Ying Liu, Luyao Geng, Weidong Zhang, et al. "Survey of Video Based Small Target Detection," Journal of Image and Graphics, Vol. 9, No. 4, pp. 122-134, December 2021. doi: 10.18178/joig.9.4.122-134.

[2] Huilan Luo, Pu Yuan, Kang Tong. A review of salient object detection methods based on Deep Learning[J]. Acta Electronica Sinica, 2021, 49(7): 1417-1427. (in Chinese)

[3] Liu J J, Hou Q, Cheng M M, et al. A Simple Pooling-Based Design for Realtime Salient Object Detection[C] //Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR), Long Beach, CA, USA ,2019: 3917-3926.

[4] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional Block Attention Module[C] //European Conference on Computer Vision (ECCV), Munich, Germany, 2018: 3-19.

[5] Lee C Y, Xie S, Gallagher P, et al. Deeply-Supervised Nets[J]. Eprint Arxiv: 1409-5185, 2014.

[6] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition[C] //IEEE Conference on Computer Vision and Pattern Recognition(CVPR), Las Vegas, NV, USA 2016: 770-778.

[7] Xie S, Tu Z. Holistically-Nested Edge Detection[J]. International Journal of Computer Vision, 2015, 125(1-3): 3-18.

[8] Zhang P, Dong W, Lu H, et al. Learning Uncertain Convolutional Features for Accurate Saliency Detection[C] //IEEE International Conference on Computer Vision (ICCV), Venice, Italy 2017: 212-221.

[9] Zhang P, Wang D, Lu H, et al. Amulet: Aggregating Multi-level Convolutional Features for Salient Object Detection[C] //IEEE Computer Society. IEEE Computer Society (ICCV), Venice, Italy, 2017:202-211.

[10] Liu N, Han J, Yang M H. PiCANet: Learning Pixel-wise Contextual Attention for Saliency Detection[J]. Eprint Arxiv: 1708. 06433, 2018.

[11] Wu Z, Su L, Huang Q. Cascaded Partial Decoder for Fast and Accurate Salient Object Detection[C] //IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019:3902-3911.

[12] Qin X, Zhang Z, Huang C, et al. BASNet: Boundary-Aware Salient Object Detection[C]// IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Long Beach, CA, USA, 2019:7471-7481.

[13] Qin X, Zhang Z, Huang C, et al. U2-Net:Going Deeper with Nested U-structure for Salient Object Detection[J]. Pattern Recognition, 2020, 106: 107404.