Multi-level Features Selection Network Based on Multi-attention for Salient Object Detection

Jianyi Ren[1,2], Zheng Wang[1,2](✉), and Meijun Sun[1,3]

[1] College of Intelligence and Computing, Tianjin University, Tianjin, China
wzheng@tju.edu.cn
[2] Tianjin Key Lab of Machine Learning, Tianjin, China
[3] Tianjin Key Lab Cognit Comp and Applicat, Tianjin, China

Abstract. Both the attention mechanism and Salient Object Detection aim to locate the most obvious regions in an image. However, common algorithms focus on micro attention but neglect the similarity in the macro perspective. Besides, they also ignore the differences among multi-scale information. To tackle these problems, we propose MFS-Net that progressive select features to predict salient regions. First, we design the Pyramid Attention module that integrates channel and spatial attention to extract semantic information for multi-scale high-level features and design the Self-Interaction Attention module to extract detailed information for multi-scale low-level features. Besides, to refine the saliency edge, we propose the Semantic-Detail Attention module which exploits high-level features to guide low-level features in a macro-attention manner. Finally, we selectively integrate global context information by the Interaction-Fusion Attention module, aiming to learn the relationship among different salient regions and alleviate the dilution effect of features. Experimental results on six benchmark datasets demonstrate that the proposed method performs well compared with 20 state-of-the-art methods.

Keywords: Salient object detection · Attention mechanism · Multi-scale features

1 Introduction

Salient Object Detection aims to locate the most obvious regions in an image. As a preprocessing step, it has been widely applied in various computer vision fields, such as object detection [18] and semantic segmentation [16]. Earlier SOD algorithms mainly used traditional methods to generate saliency maps, which rely on heuristic priors. However, these hand-crafted features are of great difficulty to capture the semantic information contained in images, thus they fail to yield satisfactory results for images with complex scenarios.

Recently, with the development of deep learning, salient object detection has made prominent progress. Due to the powerful capability to extract low-level information and high-level information at the same time [28], CNNs have

Y. Peng et al. (Eds.): ICIG 2021, LNCS 12888, pp. 315–326, 2021.
https://doi.org/10.1007/978-3-030-87355-4_27

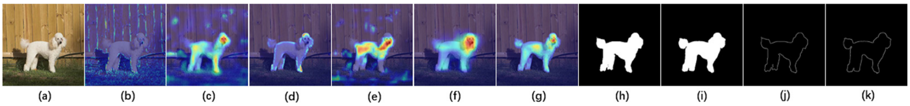

Fig. 1. Motivating examples for the proposed MFS-Net. (a) Image. (b) Low-level feature. (c) High-level feature. (d) Features extracted after SIA. (e) Features extracted after PA. (f) Features extracted after SDA. (g) Features extracted after IFA. (h) Saliency result. (i) Groung truth. (j) The boundary of (i). (k) The boundary of (h).

emerged as an important trend for SOD. Besides, the attention mechanism plays a more and more crucial role in saliency object detection. The goal of the SOD is to allocate attention to the most obvious regions in the image, and the attention mechanism focuses on local saliency information. From this perspective, the attention mechanism provides a feasible solution for SOD. Wang et al. [13] integrate the attention mechanism into the saliency object detection, which has achieved superior performance. Zhang et al. [30] propose an attention-guided model to promote the wide application of attention mechanisms in salient object detection, in which the attention mechanism makes a critical difference.

Despite CNNs and attention mechanisms have achieved excellent performance in salient object detection, there are still the following major challenges: (1) In most cases, the attention mechanism uses micro attention to generate features, such as channel and spatial attention [30]. For salient object detection which is essentially similar to the attention mechanism, they neglect the similarity in the macro perspective. (2) Many saliency studies have revealed that multi-scale features are essential for SOD [28]. Specifically, high-level features have rich semantic information (Fig. 1 (c)), and low-level features contain abundant details (Fig. 1 (b)). However, previous methods simply fuse them which ignores their different contributions. (3) Early algorithms comprehensively utilize contextual information to extract features [7,31], whereas not all contextual information contributes to the final saliency mapping. As a result, the feature extracted is incomplete while background noise is integrated.

To deal with the problem, we propose MFS-Net that selects features at multiply levels. MFS-Net emphasizes the combination of micro and macro attention mechanisms to generate saliency maps in a supervised way. In order to extract high-level semantic features and low-level detail features, we respectively propose a Pyramid Attention module (PA) and a Self-Interaction Attention module (SIA) taking advantage of the micro attention mechanism, such as the spatial and channel attention mechanism. Considering the different contributions of multi-scale features, we propose the Semantic-Detail Attention module (SDA). This module employs the macro attention mechanism to encourage high-level features to guide low-level features to suppress the background response of the original features (Fig. 1 (f)). Besides, the Interaction-Fusion Attention module (IFA) inherits the feature-enhancing ability of attention mechanisms from a macro perspective and fuse multi-scale global context information to avoid irrelevant noise caused by traditional fusion methods (Fig. 1 (g)).

In short, our contributions can be summarized as follows:

(1) In order to achieve salient object detection, we propose MFS-Net including PA, SIA, SDA and IFA modules. The PA module and the SIA module utilize the attention mechanism to extract high-level and low-level features respectively. Then the extracted semantic information guides the detailed information to suppress irrelevant background by the SDA module. Finally, the IFA module selectively integrates global context information to improve the integrity of the saliency map.
(2) Compared with 20 start-of-art SOD methods on 6 public benchmark datasets, the proposed method MFS-Net achieves remarkable performance in both quantitative and qualitative evaluation.

2 Related Works

In this section, we introduce related works from two aspects. Firstly, we review several representative salient object detection methods, and then we describe the application of attention mechanisms in various visual fields.

2.1 Salient Object Detection

Earlier saliency methods are mainly based on hand-crafted priors to estimate saliency objects, such as color contrast [3], background prior [20]. In recent years, due to the CNNs-based saliency models allow flexible feature utilization and equip powerful end-to-end capabilities, deep learning has emerged as a promising alternative for SOD.

Zhao et al. [31] proposed to use fully connected CNN to integrate global context information for saliency detection; Li et al. [7] extract multi-scale information from images of different resolutions to estimate saliency; Hu et al. [6] concatenate multi-layer features for saliency detection; Zhang et al. [26] build a directional message-passing model to better integrate multi-scale features.

The above researches demonstrate that the extraction of effective features plays a crucial role in generating a complete saliency map. Therefore, the proposed MFS-Net selectively integrates multi-scale information to generate low-level saliency feature maps guided by high-level semantic information.

2.2 Attention Mechanism

The essence of the attention mechanism is to locate obvious information and suppress useless information, which is mainly divided into spatial attention and channel attention. Attention mechanisms have been proven to be beneficial in visual tasks, such as image classification [13] and image subtitles [1].

Chen et al. [1] propose a SCA-CNN network that combines spatial and channel attention for image captioning; Li et al. [8] focus on the global context to guide target detection by using the attention mechanism; Liu et al. [10] construct

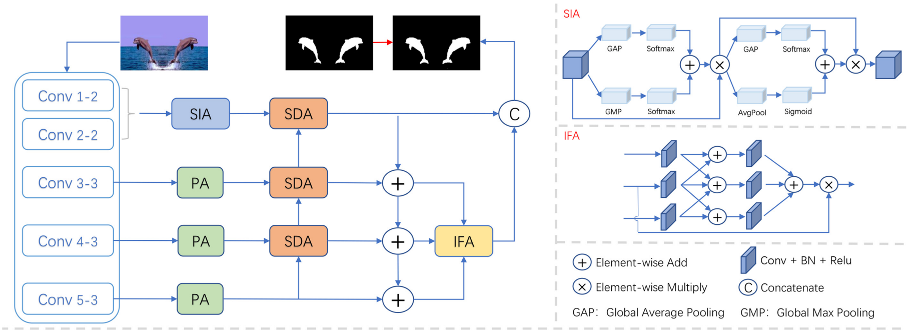

Fig. 2. Overall framework of the proposed model.

a pixel-level contextual attention model to pay attention to the information context position of each pixel; Zhang et al. [30] build a progressive attention model which sequentially generates attention features for saliency detection through the channel and spatial attention mechanisms.

The above studies demonstrate that the attention mechanism is of great help in SOD. However, most of the attention mechanisms only consider attention based on channels and spaces, named micro attention. For SOD, they ignore macro attention which is attention guidance at the feature level. On the contrary, MFS-Net proceeds with both the micro and macro aspects, which integrate global and pixel-level attention guidance, fusing the feature extraction capabilities of multi-scale information and the feature enhancement capabilities of the attention mechanism.

3 Method

In this section, we illuminate how each component made up and elucidate its effect on saliency detection. The overall architecture is illustrated in Fig. 2.

3.1 Pyramid Attention Module

In the feature extraction module, convolution operations of different levels correspond to features extraction of different levels, which directly affects the expression force of the model. For each convolutional layer containing deep semantic information, combining multi-scale information can produce more robust feature expressions. Therefore, in order to better extract the semantic information in the high-level features, we propose the Pyramid Attention module. But if we integrate multi-scale features without distinction, resulting in information redundancy and even performance degradation. Consequently, it is necessary to filtrate multi-scale information.

From Fig. 1, we can observe that the high-level feature map is rough, in which key parts may be weakened but some places that are not worthy of attention are

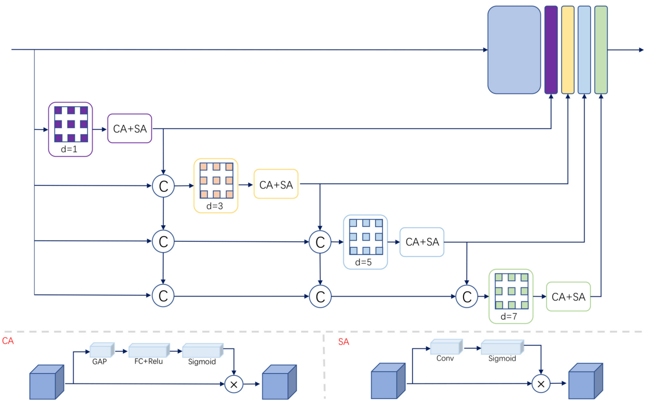

Fig. 3. Detailed structure of pyramid attention module.

given more attention. So, channel attention helps to allocate more attention to channels that show high response to salient regions. Given feature map $f^{h \times w \times c}$, we apply channel attention to generate the saliency map:

$$CA = \delta[FC(\theta(FC(GAP(f))))] \cdot f \tag{1}$$

Where GAP refers to the global average pooling layer, FC is the full connected layer, θ donates Relu function and δ represents the sigmoid operation.

The contribution of each area in the feature maps diverse a lot. We prefer to highlight salient objects rather than consider all areas equally, so we use spatial attention to pay more attention to salient regions. For increasing the receptive field and obtaining global information, we adopted the method of [4], exploiting various convolution operations whose kernels respectively are $1 \times k$, $k \times 1$ and $k \times k$ to capture features, and then apply the sigmoid function to map the feature map to $[0, 1]$ for normalization. Ultimately, the attention weight is multiplied by the original feature map to obtain the saliency map.

In the PA module, as shown in Fig. 3, given the feature map, we concatenate 4 dilated convolutional layers with dilated rates of 1, 3, 5, and 7 respectively which are combined with channel attention and spatial attention to get filtered multi-scale feature maps. Eventually, we concatenate them with the input feature map to obtain a feature map containing semantic information.

3.2 Self-interaction Attention Module

Natural images usually contain complex details. From Fig. 1, the saliency map of low-level features comprises a lot of details, some of which are beneficial for SOD but others are counterproductive. In order to extract the detailed information thoroughly from the low-level features, we propose the Self-Interaction Attention module.

In the SIA module, the score of each pixel is obtained by comparing with all other positions. Specifically, for the shallow feature $f_l^{h \times w \times c}$, it is necessary to highlight those channels which focus on foreground information and suppress other channels with background noise since each channel focuses on a different feature. Each channel can be regarded as a boundary detector, so we calculate the maximum value and the average value at the same time to obtain soft attention:

$$f_s = [\sigma(GAP(f_l)) + \sigma(GMP(f_l))] \cdot f_l \tag{2}$$

Where GMP refers to the global max-pooling layer, σ donates softmax function. GMP only pays attention to the most significant part and GAP treats all pixels equally which will inevitably merge noise, so we train f_s to make a soft choice.

In addition, in order to ensure that the attention score of each pixel is calculated both locally and globally, we add two items for global and local information extraction (Fig. 2 (upper right)). The global item is the same as the structure described above where the softmax function is combined with global average pooling of the spatial average matrix. For local item, we use local average pooling to figure out the local information similarity where a 2 × 2 pooling layer is applied to obtain the attention score of each local pixel.

$$f_o = [\sigma(GAP(f_s)) + \delta(AvP(f_s))] \cdot f_s \tag{3}$$

Considering that local information should be independent of each other, we use the sigmoid function when calculating local attention.

3.3 Semantic-Detail Attention Module

Due to multiple downsampling, high-lever features have a lot of semantic information, but they lose a lot of detailed information. At the same time, the low-level features retain rich details and background noise on account of the limitation of the receptive field. In order to refine the details of semantic features and suppress the background noise of detail features, we propose the Semantic-Detail Attention module.

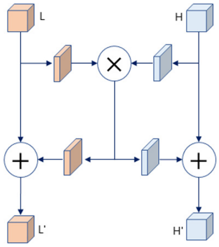

Fig. 4. Detailed structure of semantic-detail attention module.

The attention mechanism in the SDA is diverse from mentioned above. Instead of normalizing the features among [0, 1], it directly adopts the cross-guidance of attention using different feature maps on the macro level (Fig. 4). The module first applies element multiplication operation to merge the semantic feature f_h and the detailed feature f_l to extract the common feature f_c. Then the common feature f_c is combined with the previous features f_h and f_l by element addition operation respectively and the combined features $f_h^{'}$ and $f_l^{'}$ after macro attention are obtained. This attention fusion algorithm can effectively avoid the *pollution* caused by background noise. We cascade multiple SDA modules in series to make the semantic features and detailed features fully merged. Finally, the boundary of the high-level feature is sharpened and the background noise of the low-level feature is suppressed.

3.4 Interaction-Fusion Attention Module

In the process of feature transmission, some of the information will be diluted inevitably, which leads to incomplete extraction of effective features. Most methods comprehensively utilize global context information to extract features, but not all global context information contributes to the final saliency mapping. Unselected fusion will result in excessive background noise. Therefore, we propose the Interaction-Fusion Attention module.

As shown in Fig. 2 (middle right), the input of the IFA module comes from the salient feature maps after the feature filtrated at different resolutions. These feature maps are adjusted to an appropriate resolution through convolution layer, batchnorm and relu. During the process of interaction-fusion, different feature maps are fused through element addition operations to generate three feature maps that incorporate global context information. After that, we make use of the attention strategy to train three branches and obtain soft attention. Finally, for suppressing those unnecessary background noises, we utilize the macro attention mechanism to match the original features with the features filtered by the IFA module. Through the Interaction-Fusion Attention selection strategy, we can suppress those worthless background noises and integrate the filtered global context information.

3.5 Loss

We introduce the consistency enhancement loss (CEL) [12]:

$$L = \frac{|FP + FN|}{|FP + 2TP + FN|} \tag{4}$$

Where TP, FP and FN represent true-positive, false-positive and false-negative, respectively. $FP+FN$ refers to the difference between the union and intersection of the predicted map and the ground truth, while $FP + 2TP + FN$ represents the sum of the union and the intersection.

Table 1. Performance comparison with 20 methods over 6 datasets. The best three results are shown in red, blue, and green.

Methods	DUT-OMRON			PASCAL-S			DUTS-TE			ECSSD			HKU-IS			SOD		
	MAE	E-m	F-m	MAE	E-m	F-m	MAE	E-m	F-m	MAE	E-m	F-m	MAE	E-m	F-m	MAE	E-m	F-m
WSS (2017 CVPR)	0.110	0.729	0.602	0.139	0.740	0.715	0.100	0.745	0.653	0.104	0.805	0.823	0.079	0.818	0.821	0.169	0.663	0.725
SBF (2017 ICCV)	0.108	0.763	0.608	0.131	0.778	0.695	0.107	0.763	0.622	0.088	0.850	0.809	0.075	0.855	0.801	0.156	0.734	0.711
UCF (2017 ICCV)	0.120	0.760	0.621	0.115	0.811	0.726	0.112	0.775	0.631	0.069	0.890	0.844	0.062	0.886	0.823	0.164	0.742	0.695
NLDF (2017 CVPR)	0.080	0.798	0.684	0.098	0.844	0.769	0.065	0.851	0.738	0.063	0.900	0.878	0.048	0.914	0.873	0.123	0.782	0.788
AMU (2017 ICCV)	0.098	0.793	0.647	0.100	0.837	0.757	0.085	0.817	0.678	0.059	0.909	0.868	0.047	0.852	0.788	0.141	0.786	0.752
FSN (2017 ICCV)	0.066	0.844	0.706	0.093	0.853	0.766	0.066	0.861	0.729	0.053	0.924	0.872	0.044	0.928	0.858	0.126	0.809	0.772
C2S (2018 ECCV)	0.072	0.824	0.682	0.081	0.872	0.762	0.062	0.863	0.717	0.053	0.919	0.865	0.046	0.921	0.853	0.123	0.789	0.761
BDMP (2018 CVPR)	0.064	0.831	0.692	0.074	0.876	0.758	0.049	0.883	0.745	0.045	0.927	0.868	0.039	0.930	0.871	0.106	0.803	0.761
PAGRN (2018 CVPR)	0.071	0.772	0.711	0.089	0.834	0.798	0.056	0.842	0.783	0.061	0.893	0.894	0.047	0.898	0.886	0.145	0.708	0.770
PICA (2018 CVPR)	0.068	0.833	0.710	0.078	0.869	0.789	0.054	0.872	0.749	0.046	0.923	0.885	0.042	0.921	0.870	0.101	0.800	0.788
MWS (2019 CVPR)	0.109	0.729	0.609	0.133	0.735	0.713	0.091	0.743	0.684	0.096	0.791	0.840	0.084	0.787	0.814	0.166	0.660	0.734
CAPSAL (2019 CVPR)	0.104	0.669	0.563	0.075	0.871	0.810	0.062	0.846	0.743	0.082	0.843	0.819	0.055	0.885	0.843	0.147	0.698	0.688
HRS (2019 ICCV)	0.065	0.772	0.690	0.079	0.847	0.804	0.050	0.853	0.788	0.052	0.916	0.905	0.042	0.912	0.886	0.134	0.724	0.728
deepCRF (2019 ICCV)	0.057	0.838	0.738	0.082	0.852	0.790	0.059	0.854	0.744	0.049	0.921	0.896	0.039	0.925	0.881	0.121	0.776	0.785
PAGE (2019 CVPR)	0.062	0.849	0.736	0.076	0.878	0.806	0.052	0.883	0.777	0.042	0.936	0.906	0.037	0.934	0.882	0.110	0.801	0.796
TDBU (2019 CVPR)	0.061	0.867	0.739	0.071	0.883	0.775	0.048	0.892	0.766	0.041	0.937	0.880	0.038	0.933	0.878	0.104	0.821	0.767
CSNet (2020 ECCV)	0.081	0.801	0.675	0.103	0.815	0.723	0.074	0.820	0.687	0.065	0.886	0.844	0.059	0.883	0.840	0.136	0.742	0.731
SS (2020 CVPR)	0.068	0.840	0.703	0.092	0.854	0.774	0.062	0.865	0.742	0.059	0.911	0.870	0.047	0.923	0.860	0.129	0.771	0.758
EDNS (2020 ECCV)	0.076	0.811	0.682	0.094	0.837	0.790	0.065	0.851	0.735	0.068	0.894	0.872	0.046	0.918	0.873	0.142	0.754	0.776
GateNet (2020 ECCV)	0.061	0.840	0.723	0.068	0.886	0.797	0.045	0.893	0.783	0.041	0.932	0.896	0.036	0.933	0.889	-	-	-
Ours	0.067	0.845	0.718	0.072	0.885	0.805	0.048	0.897	0.790	0.043	0.936	0.897	0.037	0.939	0.881	0.102	0.815	0.794

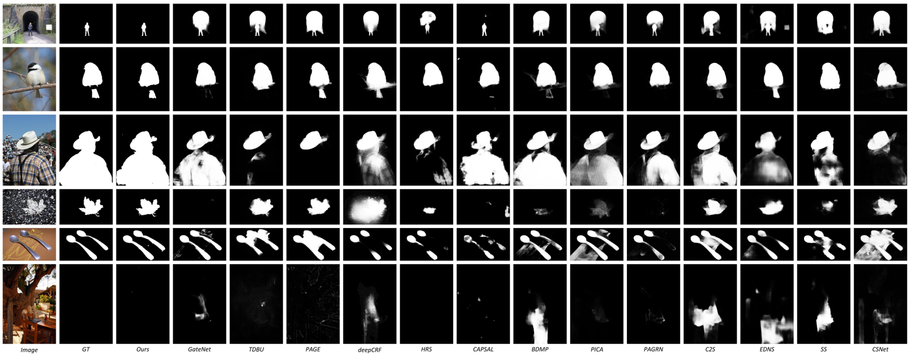

Fig. 5. Qualitative comparison of the proposed model with other methods.

4 Experiments

4.1 Datasets

We evaluate the proposed model on six public saliency detection benchmark datasets: ECSSD, DUT-OMRON, HKU-IS, PASCAL-S, DUTS and SOD, which are human-labeled with pixel-wise ground truth for quantitative evaluations.

4.2 Evaluation Criteria

To quantitatively evaluate the effectiveness of our proposed model, we adopt precision-recall (PR) curves, F-measure (Fm) score, Mean Absolute Error (MAE), and mean E-measure (Em) score as our performance measures.

4.3 Implementation Details

Following most existing state-of-the-art methods, we use DUTS-TR as our training dataset. We deploy VGG-16 trained on ImageNet as our backbone. During the training stage, we crop the image to a size of 224 × 224. Besides, we exploit random cropping and random rotation operations for data enhancement to avoid overfitting. The entire model is trained end-to-end and applies the poly strategy, where the variable is set to 0.9. To ensure model convergence, our model was trained on NVIDIA GTX 1080 Ti GPU with a batshsize of 8.

4.4 Performance Comparison

We compare the proposed MFS-Net against 20 recent SOD algorithms: WSS [14], SBF [23], UCF [29], NLDF [11], AMU [28], FSN [2], C2S [9], BDMP [26], PAGRN [30], PICA [10], MWS [22], CAPSAL [27], HRS [21], deepCRF [19], PAGE [17], TDBU [15], CSNet [5], SS [25], EDNS [24] and GateNet [32]. For fair, all the saliency maps of the above methods are provided by the authors or predicted through codes published by them.

Quantitative Comparison. In order to fully compare our proposed model with the above models, the experimental results under different metrics are listed in Table 1. It can be seen from the results that our method exhibits excellent performance, which proves the effectiveness of the proposed model. Besides, Fig. 6 shows the PR curve of the above algorithm on 6 datasets. The results reveal that our method is the most prominent in most cases, indicating that our model is highly competitive.

Qualitative Evaluation. To further illustrate the advantages of the proposed method, we provide some visual examples of different methods. Some representative examples are shown in Fig. 5. These examples reflect various scenarios, including small objects (1^{st} row), foreground disturbance (2^{nd} row), large salient object (3^{rd} row), low contrast between salient object and image background (4^{th} row), multiple salient objects (5^{th} row) and no salient object (6^{th} row). Compared with other methods, the saliency maps produced by our method are more complete and more accurate.

Table 2. Ablation study for different modules on the ECSSD dataset.

Baseline	SIA	PA	SDA	IFA	*MAE*
✓					0.071
✓	✓				0.056
✓	✓	✓			0.049
✓	✓	✓	✓		0.045
✓	✓	✓	✓	✓	0.043

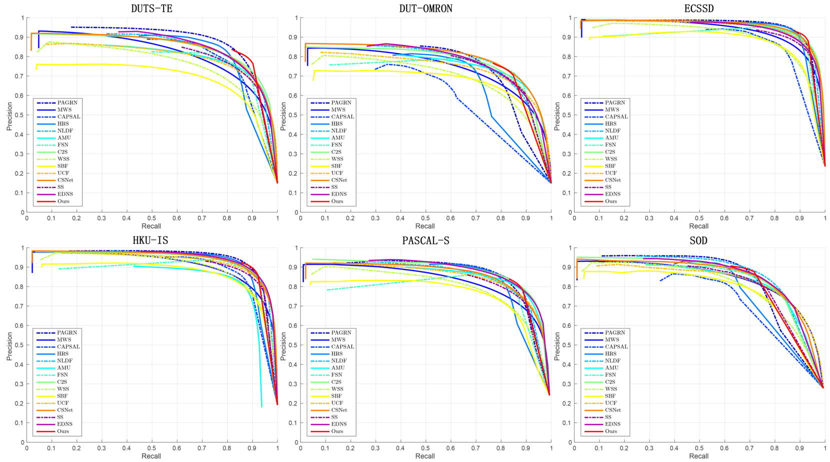

Fig. 6. Precision-Recall curves on 6 common saliency datasets.

4.5 Ablation Study

To illustrate the effectiveness of each module designed in the proposed model, we conduct the ablation study. The ablation experiments are applied on the ECSSD dataset and VGG-16 is adopted as the backbone. As shown in Table 2, the proposed model containing all components (i.e. PA, SIA, SDA and IFA) achieves the best performance, which demonstrates the necessity of each component for the proposed model to obtain the best saliency detection results.

We adopt the model which only uses high-level features after up-sampling as the baseline model, then we add each module progressively. First, we concatenate high-level features and low-feature after the SIA module which largely improves the baseline from 0.071 to 0.056 in terms of MAE. Furthermore, we add PA and get a decline of 31% in MAE compared with the basic model. On this basis, the MAE score is improved by 37% after adding SDA to both high-level features and low-level features. Finally, the combination of IFA achieves the best result.

5 Conclusion

In this paper, we propose the MFS-Net to achieve salient object detection. Taking into account the characteristics of multi-scale features, we design the PA and SIA modules to extract high-level and low-level features respectively. For refining the saliency edge, we introduce the SDA module which exploits the attention mechanism to boost detailed features guided by semantic features. Finally, we introduce the filtered global context information to alleviate the dilution effect of features. Extensive experiments on 6 datasets validate that the proposed model outperforms 20 state-of-the-art methods under different evaluation metrics.

References

1. Chen, L., et al.: SCA-CNN: spatial and channel-wise attention in convolutional networks for image captioning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5659–5667 (2017)
2. Chen, X., Zheng, A., Li, J., Lu, F.: Look, perceive and segment: finding the salient objects in images via two-stream fixation-semantic CNNs. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1050–1058 (2017)
3. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 569–582 (2014)
4. Ding, X., Guo, Y., Ding, G., Han, J.: ACNEt: strengthening the kernel skeletons for powerful cnn via asymmetric convolution blocks. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1911–1920 (2019)
5. Gao, S.H., Tan, Y.Q., Cheng, M.M., Lu, C., Chen, Y., Yan, S.: Highly efficient salient object detection with 100k parameters. arXiv preprint arXiv:2003.05643 (2020)
6. Hu, X., Zhu, L., Qin, J., Fu, C.W., Heng, P.A.: Recurrently aggregating deep features for salient object detection. In: Thirty-Second AAAI Conference on Artificial Intelligence (2018)
7. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5455–5463 (2015)
8. Li, J., et al.: Attentive contexts for object detection. IEEE Trans. Multimed. **19**(5), 944–954 (2016)
9. Li, X., Yang, F., Cheng, H., Liu, W., Shen, D.: Contour knowledge transfer for salient object detection. In: Ferrari, V., Hebert, M., Sminchisescu, C., Weiss, Y. (eds.) ECCV 2018. LNCS, vol. 11219, pp. 370–385. Springer, Cham (2018). https://doi.org/10.1007/978-3-030-01267-0_22
10. Liu, N., Han, J., Yang, M.H.: PiCANet: learning pixel-wise contextual attention for saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3089–3098 (2018)
11. Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S., Jodoin, P.M.: Non-local deep features for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6609–6617 (2017)
12. Pang, Y., Zhao, X., Zhang, L., Lu, H.: Multi-scale interactive network for salient object detection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 9413–9422 (2020)
13. Wang, F., et al.: Residual attention network for image classification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3156–3164 (2017)
14. Wang, L., eLearning to detect salient objects with image-level supervision. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 136–145 (2017)
15. Wang, W., Shen, J., Cheng, M.M., Shao, L.: An iterative and cooperative top-down and bottom-up inference network for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5968–5977 (2019)
16. Wang, W., Shen, J., Porikli, F.: Saliency-aware geodesic video object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3395–3402 (2015)

17. Wang, W., Zhao, S., Shen, J., Hoi, S.C., Borji, A.: Salient object detection with pyramid attention and salient edges. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1448–1457 (2019)
18. Xu, K., et al.: Show, attend and tell: neural image caption generation with visual attention. In: International Conference on Machine Learning, pp. 2048–2057 (2015)
19. Xu, Y., et al.: Structured modeling of joint deep feature and prediction refinement for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3789–3798 (2019)
20. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3166–3173 (2013)
21. Zeng, Y., Zhang, P., Zhang, J., Lin, Z., Lu, H.: Towards high-resolution salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 7234–7243 (2019)
22. Zeng, Y., Zhuge, Y., Lu, H., Zhang, L., Qian, M., Yu, Y.: Multi-source weak supervision for saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6074–6083 (2019)
23. Zhang, D., Han, J., Zhang, Y.: Supervision by fusion: towards unsupervised learning of deep salient object detector. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4048–4056 (2017)
24. Zhang, J., Xie, J., Barnes, N.: Learning noise-aware encoder-decoder from noisy labels by alternating back-propagation for saliency detection. arXiv preprint arXiv:2007.12211 (2020)
25. Zhang, J., Yu, X., Li, A., Song, P., Liu, B., Dai, Y.: Weakly-supervised salient object detection via scribble annotations. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 12546–12555 (2020)
26. Zhang, L., Dai, J., Lu, H., He, Y., Wang, G.: A bi-directional message passing model for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1741–1750 (2018)
27. Zhang, L., Zhang, J., Lin, Z., Lu, H., He, Y.: CapSal: leveraging captioning to boost semantics for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 6024–6033 (2019)
28. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: aggregating multi-level convolutional features for salient object detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 202–211 (2017)
29. Zhang, P., Wang, D., Lu, H., Wang, H., Yin, B.: Learning uncertain convolutional features for accurate saliency detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 212–221 (2017)
30. Zhang, X., Wang, T., Qi, J., Lu, H., Wang, G.: Progressive attention guided recurrent network for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 714–722 (2018)
31. Zhao, R., Ouyang, W., Li, H., Wang, X.: Saliency detection by multi-context deep learning. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1265–1274 (2015)
32. Zhao, X., Pang, Y., Zhang, L., Lu, H., Zhang, L.: Suppress and balance: a simple gated network for salient object detection. arXiv preprint arXiv:2007.08074 (2020)