PS-net: progressive selection network for salient object detection.

REN, J., WANG, Z. and REN, J.

2022

Noname manuscript No.
(will be inserted by the editor)

PS-Net: Progressive Selection Network for Salient Object Detection

Jianyi Ren · Zheng Wang · Jinchang Ren

Received: date / Accepted: date

Abstract Features of different scales contain distinct information. Integrating multi-scale features in an appropriate way is significant for salient object detection. However, direct concatenation or addition taken by most methods ignores the distinctions of contribution among multi-scale features. Besides, most salient object detection models fail to dynamically adjust receptive fields to fit objects of various sizes. To tackle these problems, we propose a Progressive Selection Network (PS-Net). First, we propose a Pyramid Feature Dynamic Extraction module to dynamically select appropriate receptive fields to extract high-level features by Feature Dynamic Extraction modules step by step. Besides, a Self-Interaction Attention module is designed to extract detailed information for low-level features. Finally, we design a Scale Aware Fusion module to fuse these multiple features for adequate exploitation of high-level features to refine low-level features gradually. Experimental results have demonstrated that the proposed method performs excellently in both qualitative and quantitative experiments compared with 19 methods and achieves state-of-the-art performance on four datasets.

Jianyi Ren
College of Intelligence and Computing, Tianjin University, Tianjin, China
Tianjin Key Lab of Machine Learning, Tianjin, China
E-mail: renjianyi@tju.edu.cn

Zheng Wang
College of Intelligence and Computing, Tianjin University, Tianjin, China
Tianjin Key Lab of Machine Learning, Tianjin, China
E-mail: wzheng@tju.edu.cn

Jinchang Ren
National Subsea Centre, School of Computing, Robert Gordon University, Aberdeen, UK
E-mail: Jinchang.Ren@ieee.org

Keywords salient object detection · attention mechanism · multi-scale features

1 Introduction

Salient Object Detection (SOD) aims to locate the most obvious regions in an image. As a preprocessing step, it has been widely applied in various computer vision tasks, such as object recognition [1], image editing [2], image retrieval [3], semantic segmentation [4,5] and visual tracking [6]. Earlier SOD algorithms mainly used conventional methods to generate saliency maps [7], which often rely on heuristic priors (e.g., color [8] and texture [9]). However, these hand-crafted features are of great difficulty to capture the latent semantic information in images, thus they fail to yield satisfactory results for images with complex backgrounds. Recently, with the development of deep learning, SOD has made prominent progress. Due to the powerful capability to extract low-level information and high-level information simultaneously [10,11], CNNs have emerged as an important trend for SOD, especially in complicated cases.

Despite CNNs have achieved excellent performance in SOD, there are still many challenges. (1) Many saliency studies have revealed that multi-scale features are essential for SOD [11,12,13]. Specifically, low-level features contain abundant details but full of background noise (Fig.1 (b)). On the contrary, high-level features have rich semantic information, which is helpful in locating the salient objects and suppressing the background noises (Fig.1 (c)). Therefore, it is critical to properly aggregate these features to generate satisfactory saliency maps. Existing approaches tackle the problem by integrating multiple features layer-by-layer [10,13,14]often by direct concatenation [14] or addition [15,16], which

Fig. 1 Motivating examples for the proposed PS-Net. (a) Image. (b) Original low-level feature. (c) Original high-level feature. (d) Features extracted after the PFDE module. (e) Features extracted after the SAF module. (f) Features of fusion between the SIA module and the SAF module. (g) Saliency result. (h) Ground truth.

ignores the guidance relationship using semantic information to optimize details and the differences of their contributions. (2) Besides, there is no effective extraction and utilization of multi-scale context information in every single block. When extracting high-level features, saliency objects and their surroundings are necessary to generate the final saliency maps [17]. Recently, some methods have been proposed to integrate multi-scale context information [15,18]. However, the receptive fields fail to be dynamically adjusted to fit the objects of different sizes in their methods, resulting in poor sensitivity to the change of sizes of saliency objects.

Since the attention mechanism [19,20] has been widely and successfully used for improving model performance, many networks based on the attention mechanism have been widely proposed in SOD. Therefore, to deal with the problem, we proposed a salient model PS-Net that selects features progressively at multiply levels. PS-Net emphasizes the attention mechanism to effectively integrate selected low-level appearance features and high-level semantic features to generate saliency maps in a supervised way. First, in order to extract more abundant low-level detail features, we propose a Self-Interaction Attention module (SIA) for pixel-level fusion, which fuses the global information and local information of low-level features to ensure that the attention score of each pixel is calculated both globally and locally, especially boundary-focused. Besides, due to the sizes of the salient objects that vary greatly, we propose a Pyramid Feature Dynamic Extraction module (PFDE) for the effective utilization of multi-scale context information in every single block. Different from direct concatenation or addition, the PFDE module takes advantage of the attention mechanism, named Feature Dynamic Extraction module (FDE), to dynamically adjust the receptive field in every single block to adapt to distinct sizes of salient objects (Fig.1 (d)). Finally, considering the guidance relationship using semantic information to optimize details and the different contributions between high-level features and low-level features, we propose the Scale Aware Fusion module (SAF). A spatial attention mechanism is introduced to encourage high-level features to guide low-level features and fuse them by self-learning to suppress the background response of the original features (Fig.1 (e)).

To verify the performance of PS-Net, we indicate experiment results on 6 popular SOD datasets and visualize some saliency maps. We conduct a series of ablation experiments to evaluate the effect of each module. The experiment and visual results demonstrate that PS-Net can obtain better saliency maps. We would like to highlight our contributions as follows:

(1) We introduce a Self-Interaction Attention module to extract more abundant detailed features, which ensure that the attention score of each pixel is calculated both globally and locally, especially boundary-focused.
(2) We proposed a Pyramid Feature Dynamic Extraction module to dynamically adjust the receptive field in every single block to adapt to distinct sizes of salient objects.
(3) Considering different contributions of high-level features and low-level features, we design the Scale Aware Fusion module for effective feature fusion. A spatial attention mechanism is introduced to suppress the background response of the original features.
(4) Compared with 19 start-of-the-art methods on 6 public benchmark datasets, the proposed method achieves remarkable performance in both quantitative and qualitative evaluation. We performed a lot of ablation studies, and more discussions to demonstrate the effectiveness and superiority of our proposed method.

2 Related Works

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

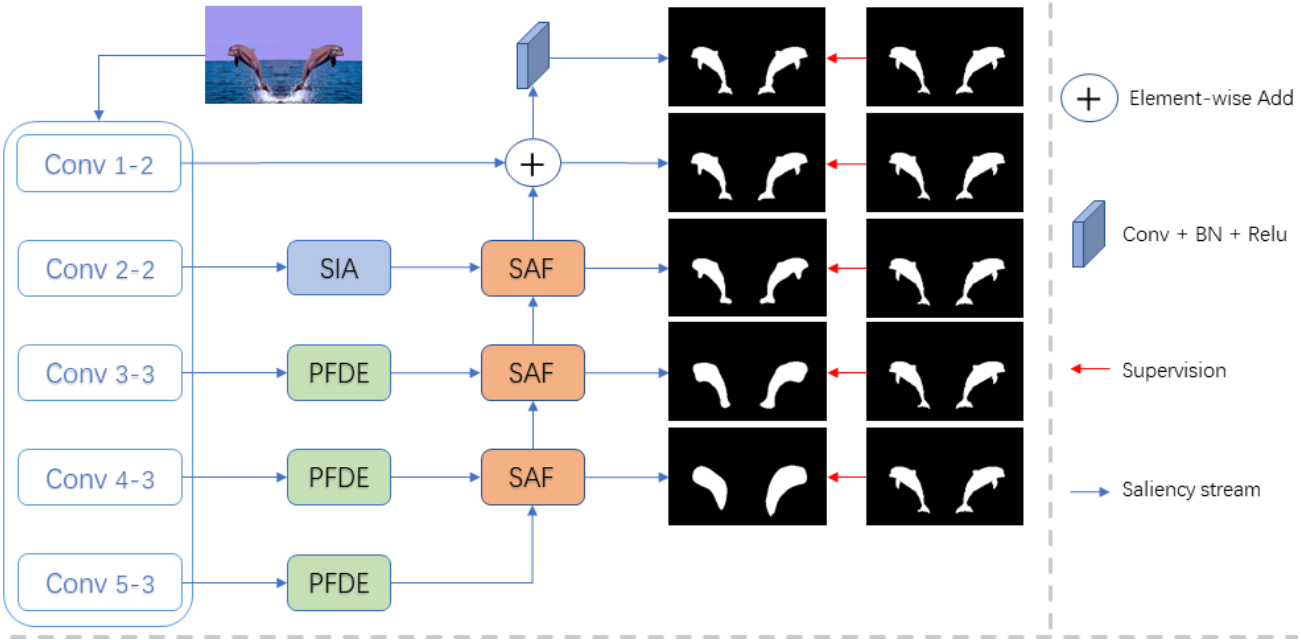

Fig. 2 The overall framework of the proposed model. Our model exploits VGG-16 blocks to extract features. The low-level features are from vgg1-2 and 2-2. The high-level features are from vgg3-3, vgg4-3 and vgg5-3. Self-Interaction Attention module(SIA) and Pyramid Feature Dynamic Extraction module(PFDE) select low-level and high-level features respectively. Scale Aware Fusion module(SAF) utilizes high-level features to guide low-level features to generate the salient map surprised by the ground truth.

2.1 Salient Object Detection

Early-stage saliency methods are mainly based on handcrafted priors to estimate saliency objects, such as color contrast [8], local contrast [21], and background priors [9]. In recent years, deep learning has emerged as a promising alternative for SOD, due mainly to the fact that CNNs-based saliency models allow flexible feature utilization and equip powerful end-to-end capabilities.

Zhao et al. [22] proposed to use a fully connected CNN to integrate global context information for saliency detection. Li et al. [23] extract multi-scale information from images of different resolutions to estimate the saliency maps. Liu et al. [24] generate prediction maps by refining edges in low-level features. Wang et al. [25] propose a model that adds low-level detail features to predict images of different scales. Hu et al. [26] concatenate multi-layer features for saliency detection. Zhang et al. [27] build a directional message-passing model to better integrate multi-scale features. There are also some methods that construct deep network architecture to optimize saliency maps [28,29].

The above researches demonstrate that the extraction of effective features plays a crucial role in generating a complete saliency map. Therefore, we propose a saliency model PS-Net that selects features progressively at multiple levels. It 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 [30], image captioning [31], and visual question and answer [32].

Chen et al. [31] propose a SCA-CNN network that combines spatial and channel attention for image captioning. Li et al. [33] focus on the global context to guide target detection by using the attention mechanism. Liu et al. [34] construct a pixel-level contextual attention model to pay attention to the information context position of each pixel. Chen et al. [35] embed the reverse attention module in the top-down approach to predict saliency maps. Zhang et al. [36] 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, while

integrating the convolutional features, most existing methods treat multi-scale features without distinction.

On the contrary, PS-Net integrates global and pixel-level attention guidance, fusing the feature extraction capabilities of multi-scale information and the feature selection 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 of the proposed method is illustrated in Fig.2.

3.1 Pyramid Feature Dynamic Extraction module

In the feature extraction module, convolution operations of different levels correspond to features extraction of different scales, which directly affects the representation capability of the model. As discussed in the introduction, low-level features contain more detailed information whilst high-level features contain affluent semantic information. Therefore, in order to better extract the semantic information in the high-level features, we propose the Pyramid Feature Dynamic Extraction module (PFDE), inspired by Atrous Spatial Pyramid Pooling (ASPP) [37].

For each convolutional layer containing deep semantic information, combining multi-scale information can produce more robust feature representations. ASPP proposes to concatenate the feature maps generated by the dilated convolution with different rates so that the salient maps encode multi-scale information under different receptive fields without distinction, resulting in information redundancy and even performance degradation. Consequently, it is necessary to mine multi-scale information for more effective fusion.

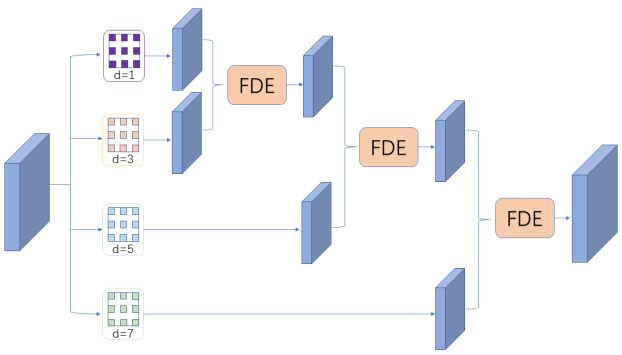

Fig. 3 Detailed structure of Pyramid Feature Dynamic Extraction (PFDE) module.

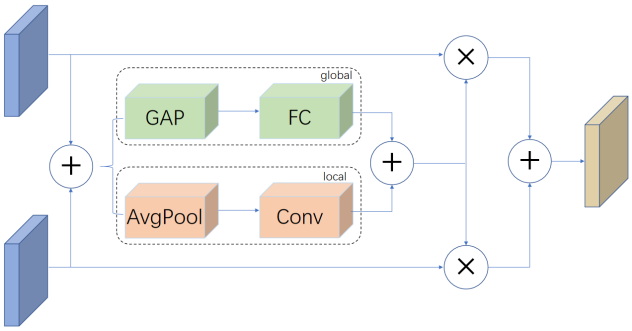

Fig. 4 The illustration of Feature Dynamic Extraction (FDE) module.

In the proposed PFDE module as shown in Fig.3, we use four parallel dilated convolutions with different dilation rates of 1, 3, 5 and 7 to capture information of different scales. After this, we design a Feature Dynamic Extraction module (FDE) to fuse differently scaled features. As shown in Fig.4, the global and local attention mechanisms are introduced to dynamically select the appropriate scale features and fuse them by self-learning. Given two features $f_1^{h\times w\times c}$ and $f_2^{h\times w\times c}$ with different reception fields, $h \times w$ represents the spatial dimension and c denotes the number of channels. First, the FDE module applies element addition operation to merge $f_1^{h\times w\times c}$ and $f_2^{h\times w\times c}$ to extract the mixed feature f_m. Then f_m locates salient objects from different receptive fields through global and local attention mechanisms respectively, which can dynamically adapt to various sizes of salient objects through self-learning. Specifically, f_m passes through the global average pooling layer and the fully connected layer followed by a repeat function respectively to obtain the global attention map f_g which has the same resolution as f_m. On the other hand, f_m goes through the average pooling layer and a convolution layer to get the local attention map f_a. Besides, the common feature f_c is combined with f_g and f_a by element addition operation respectively. Finally, the fused feature map f_f is obtained as a weighted sum as detailed below:

$$f_m = f_1 + f_2 \tag{1}$$

$$f_g = \delta(FC(GAP(f_m))) \tag{2}$$

$$f_a = \theta(conv(AvP(f_m))) \tag{3}$$

$$f_c = f_g + f_a \tag{4}$$

$$f_f = f_c \times f_1 + f_c \times f_2 \tag{5}$$

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

We employ three cascaded Feature Dynamic Extraction (FDE) modules to get the final fusion feature of four branches.

3.2 Self-Interaction Attention module

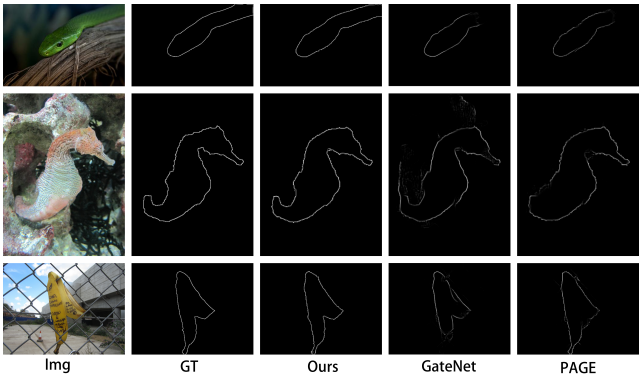

Fig. 5 Examples of boundaries of several salient objects. From left to right are original images, boundaries of Ground truth, boundaries of the proposed method, boundaries of GateNet, boundaries of PAGE.

As seen in 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. As manifested in Fig.5 where several saliency images and their corresponding boundaries are shown, the issue of unclear boundaries of saliency objects still remains a challenge, even for the latest methods with excellent performance. In order to extract the detailed information thoroughly from the low-level features and explicitly learn salient object boundaries to better locate and sharpen salient objects, 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 $f_w^{h\times w\times c}$, 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:

$$f_s = [\sigma(GAP(f_w)) + \sigma(GMP(f_w))] \times f_w \tag{6}$$

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

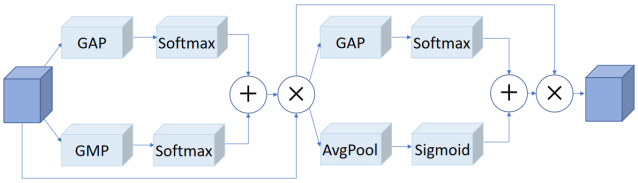

Fig. 6 Detailed structure of Self-Interaction Attention module.

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.6). 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 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.

$$f_o = [\sigma(GAP(f_s)) + \delta(AvP(f_s))] \times f_s \tag{7}$$

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

3.3 Scale Aware Fusion 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 Scale Aware Fusion (SAF) module.

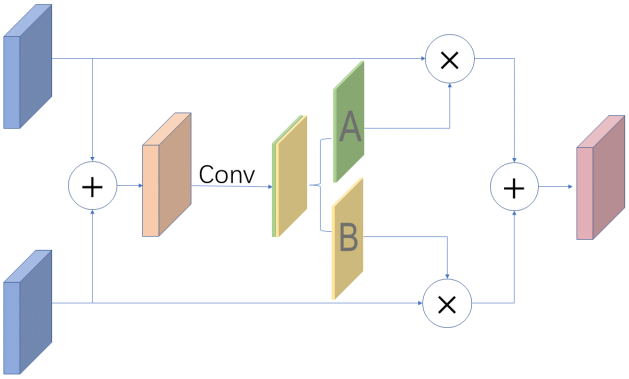

Fig. 7 Detailed structure of Scale Aware Fusion Module.

As shown in Fig.7, taking into account the attention guidance relationship and their different contributions

of multi-scale features, a spatial attention mechanism is introduced to dynamically select the appropriate scale features and fuse them. Specifically, this module first applies element addition operation to merge the semantic feature f_h and the detailed feature f_l to extract the common feature f_t. Then f_t passes through a series of convolution layers and obtains two feature maps f_A and f_B. Finally, the fused feature map f is obtained as a weighted sum:

$$f_A, f_B = D(conv(f_l + f_h)) \tag{8}$$

$$f = f_A \times f_1 + f_B \times f_h \tag{9}$$

where *conv* is cascaded by convolution, batchnorm and relu, D represents the operation of channel splitting.

This attention fusion algorithm can effectively avoid the *pollution* caused by background noise. We cascade multiple SAF modules sequentially 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 Loss

In SOD tasks, binary cross entropy loss is usually used as the loss function to evaluate the gap between the generated saliency map and the ground truth. The binary cross entropy (BCE) loss function is given as follows:

$$l = -\sum_{i=1}^{H}\sum_{j=1}^{W}[G_{ij}\log(S_{ij}) + (1 - G_{ij})\log(1 - S_{ij})] \tag{10}$$

where H, W refer to the height and width of the image respectively, G_{ij} denotes the ground truth of the pixel (i, j) and S_{ij} represents the probability of belonging to salient regions.

However, BCE cannot smoothly focus the foreground fields and treat each pixel equally, which compoundings the imbalance issue of foreground and background caused by multi-scale. To deal with the problem, two conditions need to be met: (1) It is not sensitive to changes in object size; (2) It pays more attention to the foreground field. Therefore, we introduce the consistency enhancement loss (CEL) [38]:

$$L = \frac{|FP + FN|}{|FP + 2TP + FN|} \tag{11}$$

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.

4 Experiments

4.1 Datasets

We evaluate the proposed model on six public saliency detection benchmark datasets: ECSSD [39], DUT-OMRON [9], HKU-IS [23], PASCAL-S [40], DUTS [41] and SOD [42], which are human-labeled with pixel-wise ground truth for quantitative evaluations. DUTS is currently the largest SOD dataset, including 10553 training images (DUTS-TR) and 5019 test images (DUTS-TE). DUT-OMRON contains 5168 images of complex backgrounds and high content variety. ECSSD consists of 1,000 natural-looking pictures with complex content. HKU-IS is composed of 4447 challenging images with multiple disconnected salient objects. PASCAL-S includes 850 challenging pictures. SOD contains 300 images with complex backgrounds and multiple foreground objects.

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.

MAE: defined as the average pixel-wise absolute difference between the prediction map and the ground truth.

$$MAE = \frac{1}{H \times W}\sum_{i=1}^{H}\sum_{j=1}^{W}|P(i, j) - G(i, j)| \tag{12}$$

where P refers to the predicted salient map and G denotes the ground truth.

F-measure: a comprehensive evaluation criterion calculated by a weighted combination of precision and recall.

$$F_\beta = \frac{(1 + \beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{13}$$

E-measure: combining the local pixel value with the global mean to evaluate the similarity between the predicted map and the ground truth.

Precision-Recall (PR) curve: under different thresholds, the precision and recall values can be obtained by using the predicted map and the ground truth, the thresholds are from 0 to 255.

Table 1 Performance comparison with 19 state-of-the-art methods over 6 datasets. MAE (smaller is better), mean E-measure (E-m, larger is better) and F-measure (F-m, larger is better) are used to measure the model performance. 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
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
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
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
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
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
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
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
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
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
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
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
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
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
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	0.062	0.848	0.735	0.072	0.888	0.809	0.045	0.901	0.810	0.041	0.937	0.907	0.035	0.941	0.893	0.099	0.815	0.799

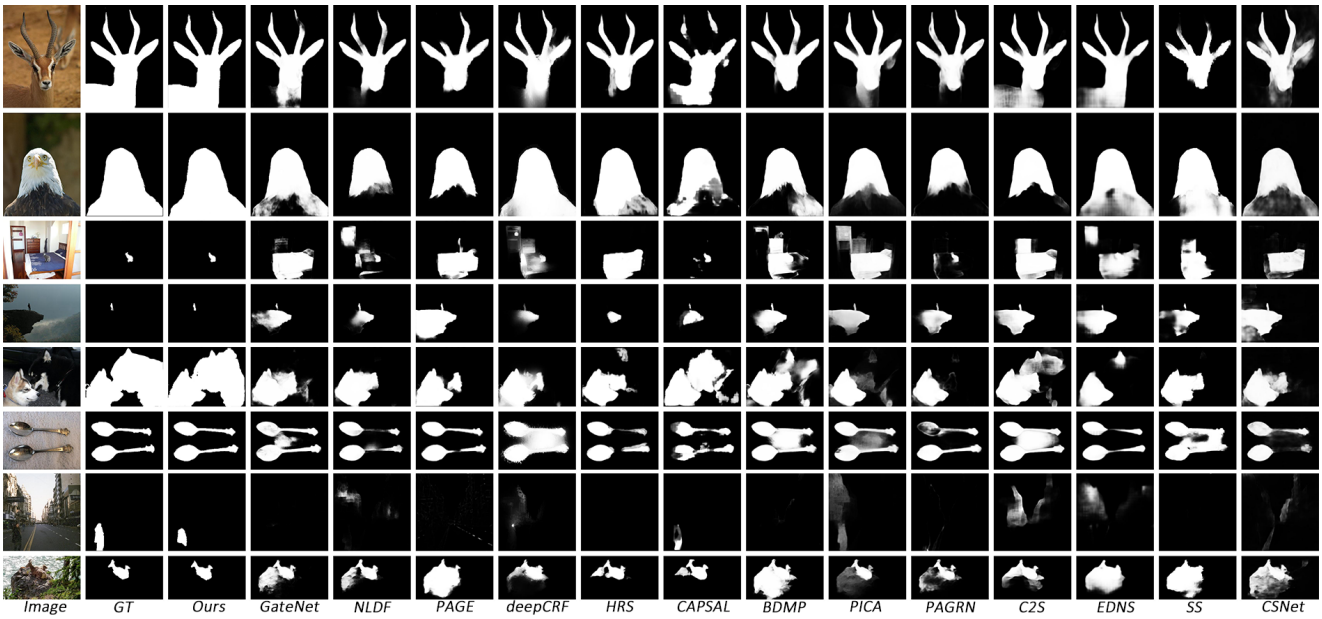

Fig. 8 Qualitative comparison of the proposed model with other state-of-the-art methods. Obviously, saliency maps produced by our model are clearer and more accurate than others and our results are more consistent with the ground truths.

4.3 Implementation Details

Following most existing state-of-the-art methods [36, 34, 25, 27], we use DUTS-TR as our training dataset. We exclude those methods which use other datasets for training, such as RADF [26] and RAS [35] which apply MASA-10K [8] for training. 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 over-fitting. The model 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. Besides, we adopted a two-step training strategy to train different components separately. Specifically, we deploy VGG-16 trained on ImageNet as our backbone and initialize other convolution layers at random. We first freeze the backbone network to train other layers for 50 epochs with a large initial learning rate, and then we train the whole network for 50 epochs with a small initial learning rate.

4.4 Performance Comparison

We compare the proposed PS-Net against 19 recent SOD algorithms: WSS [41], SBF [43], UCF [44], NLDF [10], AMU [11], FSN [45], C2S [46], BDMP [27], PAGRN [36],

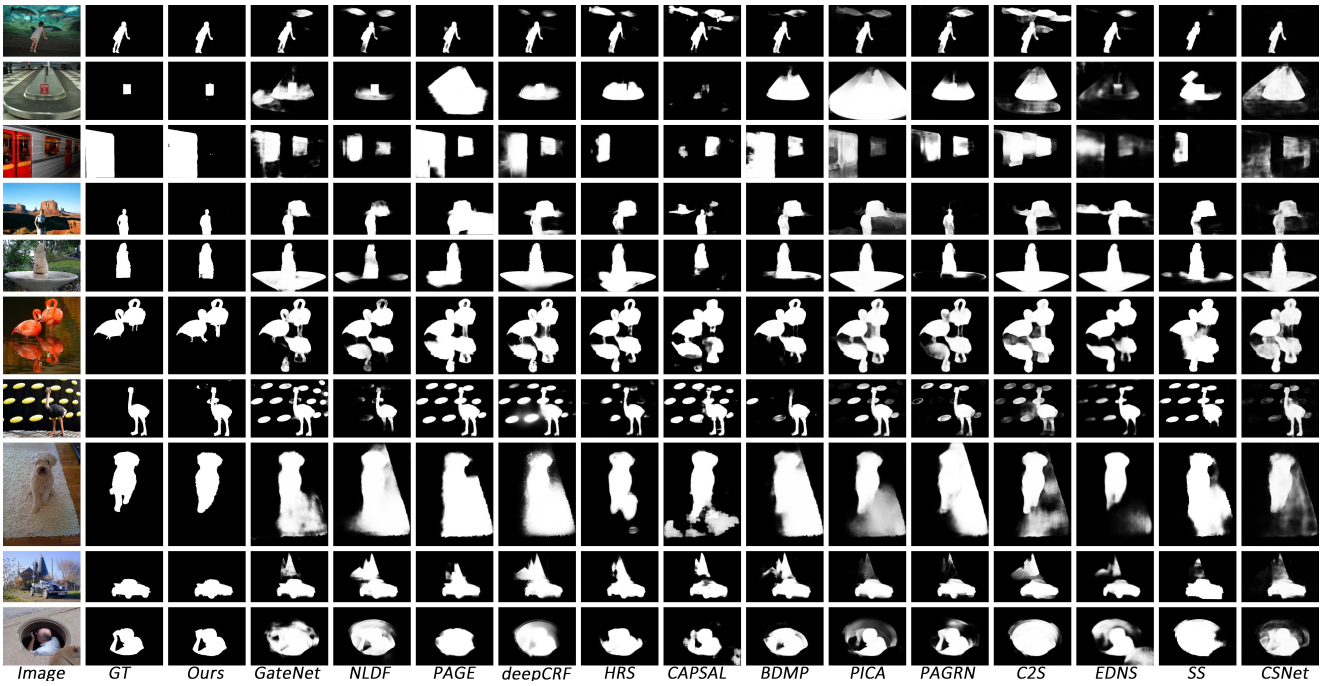

Fig. 9 Results of qualitative experiments on the salient objects with background disturbance.

PICA [34], MWS [47], CAPSAL [48], HRS [49], deepCRF [50], PAGE [51], CSNet [52], SS [53], EDNS [54] and GateNet [55]. For fair, all the saliency maps of the above methods are provided by the authors or predicted through codes published by them.

4.4.1 Quantitative Comparison

In order to fully compare our proposed model with the compared 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 validates the effectiveness of the proposed model. Besides, Fig.10 shows the PR curve of the above algorithms on the 6 datasets. The results reveal that our method is the most prominent in most cases, indicating that our model is highly competitive.

4.4.2 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.8. These examples reflect various scenarios, including large salient object (1^{st} and 2^{nd} row), small objects (3^{rd} and 4^{th} row), multiple salient objects (5^{th} and 6^{th} row), low contrast between salient object and image background (7^{th} and 8^{th} row). Compared with other methods, the saliency maps produced by our method are more complete and more accurate. Additionally, our method captures salient boundaries quite well due to its use of the Self-Interaction Attention module.

As shown in Fig.9, our method performs very well when dealing with salient objects with background disturbance due to its use of the Scale Aware Fusion Module, which takes into account the attention-guidance relationship and their different contributions.

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

Base-C	Baseline	SIA-1	SIA-12	PFDE-w	FDE	SAF	*MAE*
✓							0.064
	✓						0.071
	✓	✓					0.066
	✓	✓	✓				0.064
	✓	✓	✓	✓			0.060
	✓	✓	✓	✓	✓		0.044
	✓	✓	✓	✓	✓	✓	**0.041**

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, where VGG-16 is adopted as the backbone. As shown in Table 2, the proposed model containing all components (i.e. PFDE, SIA, and SAF) achieves the best performance, which demonstrates the necessity of each component for the proposed model to obtain the best saliency detection results.

To verify that the performance improvement of our proposed model is not caused by increasing the model

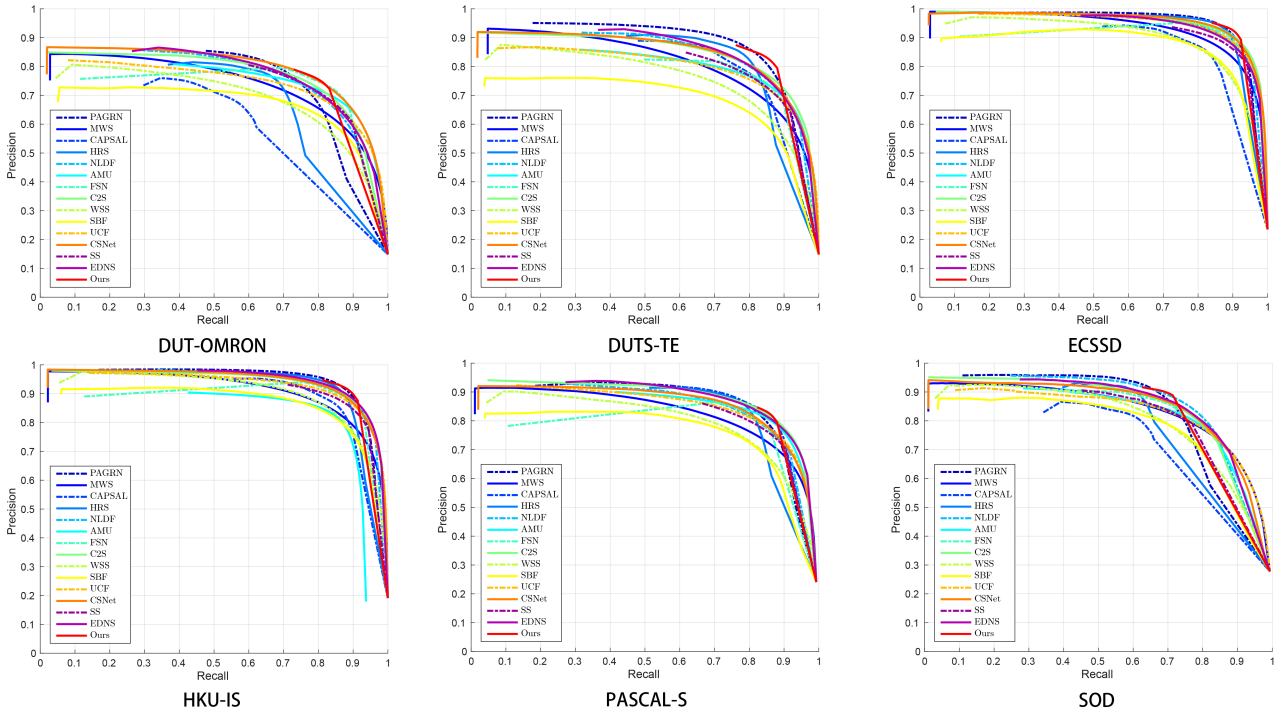

Fig. 10 Precision-Recall curves on six common saliency datasets.

complexity, we design a network based on the Baseline with similar complexity to the PS-Net by adding channels, which is called Base-C in Table 2. The experiment shows that our proposed PS-Net achieves notable improvement than the Base-C (36% in terms of MAE).

We adopt the model which only uses high-level features after up-sampling as the baseline model, then we add each module progressively. First, in order to verify the function of each part of the SIA module more accurately, we extract low-level features after the first part of the SIA module which is shown in Formula 6 and after the whole SIA module respectively. Integrating high-level features and low-level features by addition, we improve the baseline from 0.071 to 0.066 and 0.064 respectively in terms of MAE. Furthermore, we add the PFDE module where the FDE module is replaced by the addition operation, which is called PFDE-w in Table 2. The result shows that we get a decline of 15% in MAE compared with the basic model. On this basis, the MAE score is improved by 38% after adding FDE to the PFDE module. Finally, the combination of SAF achieves the best result.

5 Conclusion

In this paper, we propose a Progressive Selection Network (PS-Net) for effective salient object detection. Taking into account the characteristics of multi-scale features, we design the PFDE module to aggregate high-level features dynamically. For refining the saliency edge, we propose the SIA module to extract low-level features. Besides, considering the different contributions of high-level features and low-level features, we propose the SAF module which exploits high-level features to guide low-level features. Extensive experiments on 6 datasets validate that the proposed model outperforms 19 state-of-the-art methods under different evaluation metrics.

Acknowledgements The authors wish to acknowledge the support for the research work from the National Natural Science Foundation of China under grant Nos.[61772360] ,[61876125] and [62076180].

Conflict of interest

The authors declare that they have no conflict of interest.

References

1. U. Rutishauser, D. Walther, C. Koch, P. Perona, in *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, vol. 2 (IEEE, 2004), vol. 2, pp. II–II
2. M.M. Cheng, F.L. Zhang, N.J. Mitra, X. Huang, S.M. Hu, ACM Transactions on Graphics (TOG) **29**(4), 1 (2010)
3. J. He, J. Feng, X. Liu, T. Cheng, T.H. Lin, H. Chung, S.F. Chang, in *2012 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2012), pp. 3005–3012

4. W. Wang, J. Shen, F. Porikli, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 3395–3402
5. W. Wang, J. Shen, H. Sun, L. Shao, IEEE Transactions on Circuits and Systems for Video Technology **28**(8), 1727 (2017)
6. S. Hong, T. You, S. Kwak, B. Han, in *International conference on machine learning* (2015), pp. 597–606
7. Z. Jiang, L.S. Davis, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2013), pp. 2043–2050
8. M.M. Cheng, N.J. Mitra, X. Huang, P.H. Torr, S.M. Hu, IEEE transactions on pattern analysis and machine intelligence **37**(3), 569 (2014)
9. C. Yang, L. Zhang, H. Lu, X. Ruan, M.H. Yang, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2013), pp. 3166–3173
10. Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, P.M. Jodoin, in *Proceedings of the IEEE Conference on computer vision and pattern recognition* (2017), pp. 6609–6617
11. P. Zhang, D. Wang, H. Lu, H. Wang, X. Ruan, in *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 202–211
12. L. Wang, H. Lu, X. Ruan, M.H. Yang, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2015), pp. 3183–3192
13. Q. Hou, M.M. Cheng, X. Hu, A. Borji, Z. Tu, P.H. Torr, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 3203–3212
14. Z. Wu, L. Su, Q. Huang, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2019), pp. 3907–3916
15. J.J. Liu, Q. Hou, M.M. Cheng, J. Feng, J. Jiang, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 3917–3926
16. J.X. Zhao, J.J. Liu, D.P. Fan, Y. Cao, J. Yang, M.M. Cheng, in *Proceedings of the IEEE/CVF International Conference on Computer Vision* (2019), pp. 8779–8788
17. Y. Qin, K. Kamnitsas, S. Ancha, J. Nanavati, G. Cottrell, A. Criminisi, A. Nori, in *International Conference on Medical Image Computing and Computer-Assisted Intervention* (Springer, 2018), pp. 603–611
18. L.C. Chen, G. Papandreou, F. Schroff, H. Adam, arXiv preprint arXiv:1706.05587 (2017)
19. C. Yu, J. Wang, C. Peng, C. Gao, G. Yu, N. Sang, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 1857–1866
20. J. Hu, L. Shen, G. Sun, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2018), pp. 7132–7141
21. D.A. Klein, S. Frintrop, in *2011 International Conference on Computer Vision* (IEEE, 2011), pp. 2214–2219
22. R. Zhao, W. Ouyang, H. Li, X. Wang, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 1265–1274
23. G. Li, Y. Yu, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2015), pp. 5455–5463
24. N. Liu, J. Han, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2016), pp. 678–686
25. T. Wang, A. Borji, L. Zhang, P. Zhang, H. Lu, in *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 4019–4028
26. X. Hu, L. Zhu, J. Qin, C.W. Fu, P.A. Heng, in *Thirty-second AAAI conference on artificial intelligence* (2018)
27. L. Zhang, J. Dai, H. Lu, Y. He, G. Wang, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 1741–1750
28. J. Kuen, Z. Wang, G. Wang, in *Proceedings of the IEEE Conference on computer Vision and Pattern Recognition* (2016), pp. 3668–3677
29. H. Song, W. Wang, S. Zhao, J. Shen, K.M. Lam, in *Proceedings of the European conference on computer vision (ECCV)* (2018), pp. 715–731
30. F. Wang, M. Jiang, C. Qian, S. Yang, C. Li, H. Zhang, X. Wang, X. Tang, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 3156–3164
31. L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, T.S. Chua, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2017), pp. 5659–5667
32. H. Xu, K. Saenko, in *European Conference on Computer Vision* (Springer, 2016), pp. 451–466
33. J. Li, Y. Wei, X. Liang, J. Dong, T. Xu, J. Feng, S. Yan, IEEE Transactions on Multimedia **19**(5), 944 (2016)
34. N. Liu, J. Han, M.H. Yang, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 3089–3098
35. S. Chen, X. Tan, B. Wang, X. Hu, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 234–250
36. X. Zhang, T. Wang, J. Qi, H. Lu, G. Wang, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 714–722
37. L.C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, A.L. Yuille, IEEE transactions on pattern analysis and machine intelligence **40**(4), 834 (2017)
38. Y. Pang, X. Zhao, L. Zhang, H. Lu, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 9413–9422
39. Q. Yan, L. Xu, J. Shi, J. Jia, in *Proceedings of the IEEE conference on computer vision and pattern recognition* (2013), pp. 1155–1162
40. Y. Li, X. Hou, C. Koch, J.M. Rehg, A.L. Yuille, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2014), pp. 280–287
41. L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, X. Ruan, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2017), pp. 136–145
42. V. Movahedi, J.H. Elder, in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops* (IEEE, 2010), pp. 49–56
43. D. Zhang, J. Han, Y. Zhang, in *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 4048–4056
44. P. Zhang, D. Wang, H. Lu, H. Wang, B. Yin, in *Proceedings of the IEEE International Conference on computer vision* (2017), pp. 212–221
45. X. Chen, A. Zheng, J. Li, F. Lu, in *Proceedings of the IEEE International Conference on Computer Vision* (2017), pp. 1050–1058
46. X. Li, F. Yang, H. Cheng, W. Liu, D. Shen, in *Proceedings of the European Conference on Computer Vision (ECCV)* (2018), pp. 355–370
47. Y. Zeng, Y. Zhuge, H. Lu, L. Zhang, M. Qian, Y. Yu, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 6074–6083
48. L. Zhang, J. Zhang, Z. Lin, H. Lu, Y. He, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 6024–6033
49. Y. Zeng, P. Zhang, J. Zhang, Z. Lin, H. Lu, in *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 7234–7243

50. Y. Xu, D. Xu, X. Hong, W. Ouyang, R. Ji, M. Xu, G. Zhao, in *Proceedings of the IEEE International Conference on Computer Vision* (2019), pp. 3789–3798
51. W. Wang, S. Zhao, J. Shen, S.C. Hoi, A. Borji, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2019), pp. 1448–1457
52. S.H. Gao, Y.Q. Tan, M.M. Cheng, C. Lu, Y. Chen, S. Yan, arXiv preprint arXiv:2003.05643 (2020)
53. J. Zhang, X. Yu, A. Li, P. Song, B. Liu, Y. Dai, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2020), pp. 12,546–12,555
54. J. Zhang, J. Xie, N. Barnes, arXiv preprint arXiv:2007.12211 (2020)
55. X. Zhao, Y. Pang, L. Zhang, H. Lu, L. Zhang, arXiv preprint arXiv:2007.08074 (2020)