Pyramid Feature Attention Network for Saliency detection

Ting Zhao, Xiangqian Wu
School of Computer Science and Technology, Harbin Institute of Technology
17S003073@stu.hit.edu.cn, xqwu@hit.edu.cn✉

Abstract

Saliency detection is one of the basic challenges in computer vision. How to extract effective features is a critical point for saliency detection. Recent methods mainly adopt integrating multi-scale convolutional features indiscriminately. However, not all features are useful for saliency detection and some even cause interferences. To solve this problem, we propose Pyramid Feature Attention network to focus on effective high-level context features and low-level spatial structural features. First, we design Context-aware Pyramid Feature Extraction (CPFE) module for multi-scale high-level feature maps to capture rich context features. Second, we adopt channel-wise attention (CA) after CPFE feature maps and spatial attention (SA) after low-level feature maps, then fuse outputs of CA & SA together. Finally, we propose an edge preservation loss to guide network to learn more detailed information in boundary localization. Extensive evaluations on five benchmark datasets demonstrate that the proposed method outperforms the state-of-the-art approaches under different evaluation metrics.

1. Introduction

Saliency detection aims to locate the important parts of natural images which attract our attention. As the preprocessing of computer vision applications, e.g. object detection[8, 35], visual tracking[1, 14], image retrieval[10, 13] and semantic segmentation[9], saliency detection attracts many researchers. Currently, the most effective saliency detection methods are based on the fully convolutional network (FCN). FCN stacks multiple convolution and pooling layers to gradually increase the receptive field and generate the high-level semantic information, which plays a crucial role in saliency detection. However, the pooling layers reduce the size of the feature maps and deteriorate the boundaries of the salient objects.

To deal with this problem, some works introduce hand-craft features to preserve the boundaries of salient objects[18, 28]. [18] extracts the hand-craft features to compute the salient values of super-pixels. [28] partitions the image into regions by hand-craft features. When generating saliency maps, the hand-craft features and the CNN high-level features are complementary but extracted separately in these methods. However, it is difficult to effectively fuse the complementary features extracted separately. Furthermore, hand-craft features extraction is a time-consuming procedure.

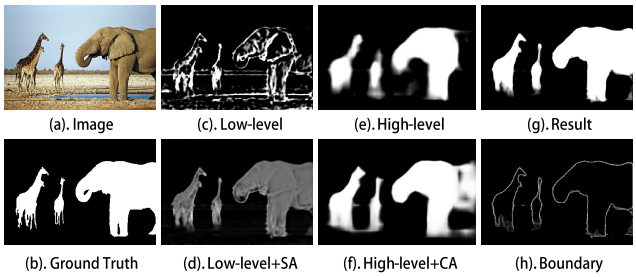

Figure 1. An example of applying Pyramid Feature Attention network.(a) and (b) represent the input image and corresponding Ground Truth. (c) and (d) mean low-level features without or with spacial attention. (e) and (f) are high-level features without or with channel-wise attention. (g) and (h) represent the results from our method and the boundary map of (g) generated by Laplace operator.

Besides hand-craft features, some works discover that the features from different layers of the network are also complementary and integrate the multi-scale features for saliency detection [15, 43, 29]. More specifically, the features at deep layers typically contain global context-aware information, which are suitable to locate the salient regions correctly. The features at shallow layers contain the spatial structural details, which are suitable to locate boundaries. These methods fused different scale features without considering their different contribution for saliency, it is not optimal for saliency detection. To overcome these problems, attention model [45] and gate function [42] are introduced to the saliency detection networks. However, the methods ignore the different characteristics of the high-level and low-level features, which may affect the extraction of effective features.

In this paper, we propose a novel salient object detection method named Pyramid Feature Attention (PFA) network. In consideration of the different characteristics of different level features (Fig.1 (c,e)), the saliency maps from low-level features contain many noises, while the saliency maps from high-level features only get an approximate area. Therefore, for high-level features, inspired by SIFT[23] feature extraction algorithm, we design a context-aware pyramid feature extraction(CPFE) module to get multi-scale multi-receptive-field high-level features, and then we use channel-wise attention(CA) to select appropriate scale and receptive-field for generating saliency regions. In training process, CA assigns large weights to the channels which play important role for saliency detection(Fig.1 (f)). In order to refine the boundaries of saliency regions, we fuse low-level features with edge information. But not all edge information is effective for refining saliency maps, we expect to focus on the boundaries between salient objects and background. So we use spatial attention to better focus on the effective low-level features, and obtain clear saliency boundaries(Fig.1 (d)). After the processing of different attention mechanisms, the high-level features and low-level features are complementary-aware and suitable to generate saliency map. In addition, different from previous saliency detection approaches, we propose an edge preservation loss to guide network to learn more detailed information in boundary localization. With the above considerations, the proposed method PFA network can produce good saliency maps.

In short, our contributions are summarized as follows:

1. We propose a Pyramid Feature Attention (PFA) network for image saliency detection. For high-level feature, we adopt a context-aware pyramid feature extraction module and a channel-wise attention module to capture rich context information. For low-level feature, we adopt spatial attention module to filter out some background details.

2. We design a novel edge preservation loss to guide network to learn more detailed information in boundary localization.

3. The proposed model achieves the state-of-the-art on several challenging datasets. The experiments prove the effectiveness and superiority of the proposed method.

2. Related Works

2.1. Salient Object Detection

In the past decade, there are a number of approaches for saliency detection. Early approaches[5, 38, 39, 17] estimate the salient value based on hand-crafted features. Those approaches detect salient objects with humanlike intuitive feelings and heuristic priors, such as color contrast[5], boundary background[38, 39] and center prior[17]. These direct techniques are known to be friendly to keep fine image structure. Nevertheless, the hand-craft features and priors can hardly capture high-level and global semantic knowledge about the objects.

In recent years, many efforts about various network architectures have been made in saliency detection. Some experiments[15, 18, 29] show that high-level features in deep layers encode the semantic information for getting an abstract description of objects, while low-level features in shallow layers keep spatial details for reconstructing the object boundaries (Fig.1 (c,e)). Accordingly, some works bring multi-level features into saliency detection. Hou et al. [15] propose a saliency method by introducing short connections to the skip-layer structures within the HED architecture. Wang et al. [31] propose a saliency detection method based on recurrent fully convolutional networks (RFCNs). Luo et al. [24] combine the local and global information through a multi-resolution grid structure. Zhang et al. [43] aggregate multi-level features by concatenating feature maps from both high levels and low levels directly. Zhang et al. [42] propose a bi-directional message passing module, where messages can transmit mutually controlled by gate function. However, some features may cause interferences in saliency detection. How to get various features and select effective ones becomes an important problem in saliency detection.

2.2. Attention Mechanisms

Attention mechanisms have been successfully applied in various tasks such as machine translation [11], object recognition [25], image captioning [3, 36], visual question answering [34, 41] and pose estimation [6]. Chu et al. [6] propose a network with multi-context attention mechanism into an end-to-end framework for human pose estimation. Chen et al. [3] propose a SCA-CNN network that incorporates spatial and channel-wise attention in CNN for image captioning. Zhang et al.[45] propose a progressive attention guided network which generates attentive features by channel-wise and spatial attention mechanisms sequentially for saliency detection.

Due to attention mechanisms have great ability to select features, it is a perfect fit for saliency detection. While integrating the convolutional features, most existing methods treat multi-level features without distinction. Some methods adopted certain valid strategies, such as gate function[42] and progressive attention[45], but those methods select features in a certain direction and ignore the differences between high-level and low-level features. Different from them, for high-level feature, we adopt context-aware pyramid feature extraction(CPFE) module and channel-wise attention module to capture rich context information. In CPFE module, we adopt multi-scale atrous convolutions on the side of three high-level blocks of VGG net, then channel-wise attention mechanism assigns large

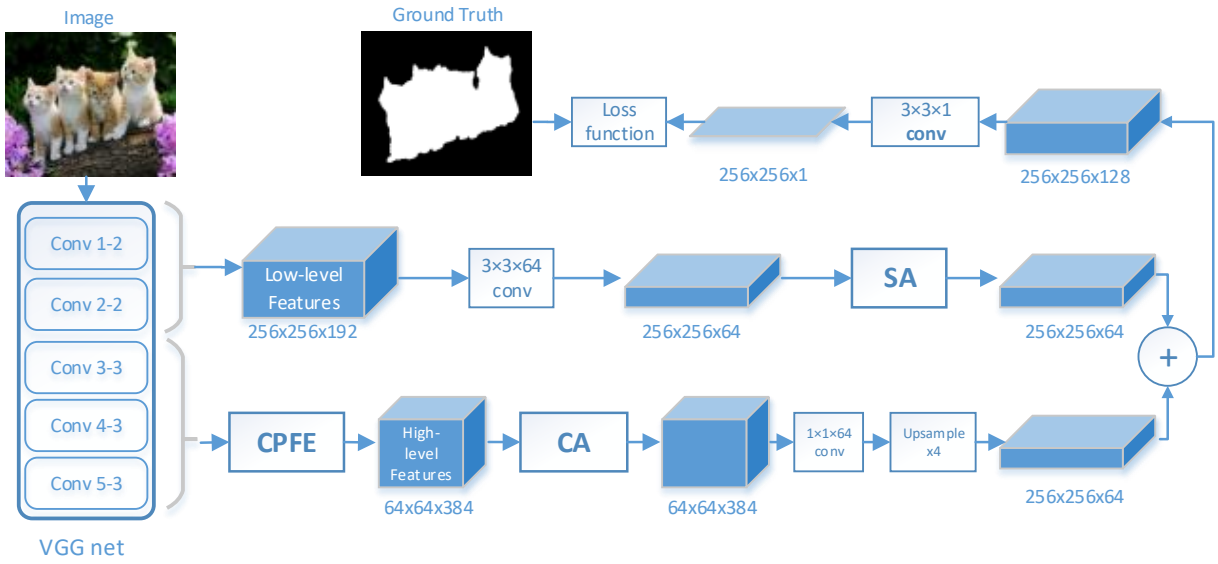

Figure 2. The overall architecture of our method. CPFE means context-aware pyramid feature extraction. The high-level features are from vgg3-3, vgg4-3 and vgg5-3. The low-level features are from vgg1-2 and 2-2, which upsample to the size of vgg1-2.

weights to channels which show high response to salient objects. For low-level feature, there exists some background regions which distract the generation of saliency map. Spatial attention mechanism filters out some background details according to high-level features and focus more on the foreground regions, which helps to generate effective features for saliency prediction.

3. Pyramid Feature Attention Network

In this paper, we propose a novel saliency detection method, which contains a context-aware pyramid feature extraction module and a channel-wise attention module to capture context-aware multi-scale multi-receptive-field high-level features, a spatial attention module for low-level feature maps to refine salient object details and an effective edge preservation loss to guide network to learn more detailed information in boundary localization. The overall architecture is illustrated in Fig.2.

3.1. Context-aware pyramid feature extraction

Visual context is quite important for saliency detection. Existing CNN models learn features of objects by stacking multiple convolutional and pooling layers. However, the salient objects have large variations in scale, shape and position. Previous methods usually directly use the bottom-to-up convolution and pooling layers, that may not be effectively to handle these complicated variations. Inspired by the feature extraction of SIFT[23], we try to design a novel module to extract the features of scale, shape and location invariances. The scale-invariant feature transform (SIFT) is a feature detection algorithm in computer vision to detect and describe local features in images. The algorithm proposed the Laplassian of Gaussian representation[23] which fused scale space representations and pyramid multi-resolution representations. The scale space representations which are processed by several different Gaussian kernel functions with same resolution; and the pyramid multi-resolution representations which are processed by down samples with different resolutions. Similar with Gaussian function in SIFT, we use atrous convolution [4] to get features with same scale but different receptive fields. Similar with pyramid multi-resolution representations in SIFT, we take conv3-3, conv4-3 and conv5-3 of VGG-16 [27] to extract multi-scale features.

Specifically, the context-aware pyramid feature extraction module is shown in Fig.3. We take conv 3-3 , conv 4-3 and conv 5-3 in VGG-16 as the basic high-level features. To make the final extracted high-level features contain the features of scale and shape invariances, we adopt atrous convolution with different dilation rates, which are set to 3, 5 and 7 to capture multi-receptive-field context information. Then we combine the feature maps from different atrous convolutional layers and a 1×1 dimension reduction feature by cross-channel concatenation. After this, we get three different scale features with context-aware information, we upsample the two smaller ones to the largest one. Finally, we combine them by cross-channel concatenation as the output of the context-aware pyramid feature extraction module.

3.2. Attention mechanism

We exploit context-aware pyramid feature extraction to get multi-scale multi-receptive-field high-level features. Different features have different semantic values to generate saliency maps. But most existing methods integrate multi-

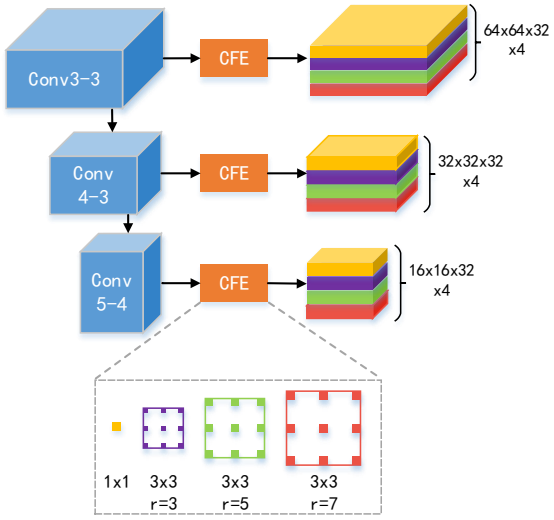

Figure 3. Detailed structure of context-aware pyramid feature extraction. A context-aware feature extraction module takes a feature from a side output of net as input and it contains three 3×3 convolutional layers with different dilation rates and a 1×1 convolutional layers, the output channel of each convolutional layer is 32.

scale features without distinction, which lead to information redundancy. More importantly, inaccurate information at some levels would lead to a performance degradation or even wrong prediction. It is significant to filter these features and fucus more on valuable features. In this subsection we will talk about the attention mechanisms in PFA network. According to the characteristics of different level features, we adopt channel-wise attention for high-level features and spacial attention for low-level features to select effective features. In addition, we don't use spacial attention for high-level features, because high-level features contain high abstract semantics[16, 45], there is no need to filter spacial information. While, we don't use channel-wise attention for low-level feature, because there are almost no semantic distinctions among different channels of low-level features.

3.2.1 Channel-wise attention

Different channels of features in CNNs generate response to different semantics[16]. From Fig.1, the saliency map from high-level features is just a rough result, some essential regions may be weakened. We add channel-wise attention (CA) [16, 3] module after context-aware pyramid feature extraction to weighted multi-scale multi-receptive-field high-level features. The CA will assign larger weight to channels which show high response to salient objects .

We unfold high-level features $\boldsymbol{f}^h \in \mathbb{R}^{W\times H\times C}$ as $\boldsymbol{f}^h$= [$\boldsymbol{f}_1^h, \boldsymbol{f}_2^h, ..., \boldsymbol{f}_C^h$], where $\boldsymbol{f}_i^h \in \mathbb{R}^{W\times H}$ is the *i*-th slice of $\boldsymbol{f}^h$ and C is the total channel number. First, we apply average pooling to each $\boldsymbol{f}_i^h$ to obtain a channel-wise feature vector $\boldsymbol{v}^h \in \mathbb{R}^C$. After that, two consecutive fully connected(FC) layer to fully capture channel-wise dependencies(see Fig.4). As [16], to limit model complexity and aid generalisation, we encode the channel-wise feature vector by forming a bottleneck with two FC layers around the non-linearity. Then, through using sigmoid operation, we take the normalization processing measures to the encoded channel-wise feature vector mapped to [0,1].

$$CA = F(\boldsymbol{v}^h, W) = \sigma_1(fc_2(\delta(fc_1(\boldsymbol{v}^h, W_1)), W_2)) \quad (1)$$

Where W refers to parameters in channel-wise attention block, σ_1 refers to sigmoid operation, fc refers to FC layers, δ refers to the ReLU function. The final output $\widetilde{f}^h$ of the block is obtained by weighting the context-aware pyramid features with CA.

$$\widetilde{f}^h = CA \cdot f^h \quad (2)$$

3.2.2 Spacial attention

Natural images usually contains a wealth of details of foreground and complex background. From Fig.1, the saliency map from low-level features contains a lot of details which easily brings bad results. In saliency detection, we want to obtain detailed boundaries between salient objects and background without other texture which can distract human attention. Therefore, instead of considering all spatial positions equally, we adopt spatial attention to focus more on the foreground regions, which helps to generate effective features for saliency prediction.

We represent low-level features as $\boldsymbol{f}^l \in \mathbb{R}^{W\times H\times C}$. The set of spatial locations is denoted by $\mathbb{R} = \{(x, y)|x = 1, ..., W; y = 1, ..., H\}$, where *j* =(x,y) is the spatial coordinate of low-level features. For increasing receptive field and getting global information but not increasing parameters, similar to [26], we apply two convolution layers ,one's kernel is 1×k and the other's is k×1, for high-level feature to capture spacial concerns(see Fig.4). Then, using sigmoid operation, we take the normalization processing measures to the encoded spacial feature map mapped to [0,1].

$$C_1 = conv_2(conv_1(\widetilde{f}^h, W_1^1), W_1^2)) \quad (3)$$

$$C_2 = conv_1(conv_2(\widetilde{f}^h, W_2^1), W_2^2)) \quad (4)$$

$$SA = F(\widetilde{f}^h, W) = \sigma_2(C_1 + C_2) \quad (5)$$

Where W refers to parameters in spacial attention block, σ_2 refers to sigmoid operation, $conv_1$ and $conv_2$ refers to

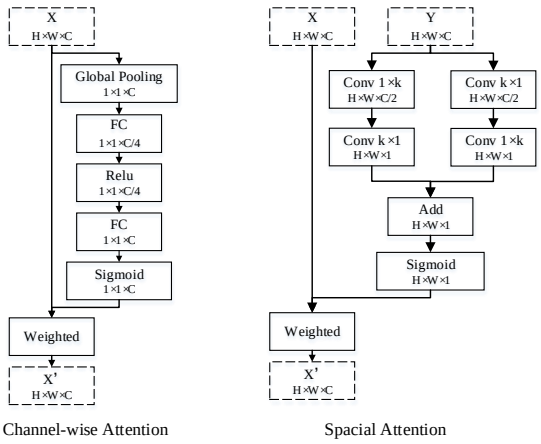

Figure 4. Channel-wise attention (left) and spacial attention (right). Where X and X' mean weighted feature and weighting feature respectively, Y means context-aware high-level feature after CA in this paper.

1×k×C and k×1×1 convlution layer respectively and we set k=9 in experiment. The final output $\widetilde{f}^l$ of the block is obtained by weighting f^l with SA.

$$\widetilde{f}^l = SA \cdot f^l \tag{6}$$

3.3. Loss function

In machine learning and mathematical optimization, loss functions represent the price paid for inaccuracy of predictions in classification problems. In saliency object detection, we always use the cross-entropy loss between the final saliency map and the ground truth. The loss function is defined as:

$$\begin{aligned} L_S = -\sum_{i=0}^{size(Y)} (\alpha_s Y_i log(P_i) \\ +(1-\alpha_s)(1-Y_i)log(1-P_i)) \end{aligned} \tag{7}$$

where Y means the ground truth and P means the saliency map of network output, α_s means a balance parameter of positive and negative samples and we set $\alpha_s = 0.528$ which calculated from groundtruth of the training set. However, the loss function just provides general guidance to generate saliency map. We use a simpler strategy to emphasize generation of the salient object boundaries details. First, we use Laplace Operator[12] to get boundaries of ground truth and saliency map of network output, and then we use the cross-entropy loss to supervise the generation of salient object boundaries.

$$\Delta f = \frac{\partial^2 f}{\partial x^2} + \frac{\partial^2 f}{\partial y^2} \tag{8}$$

$$\Delta \widetilde{f} = abs(tanh(conv(f, K_{laplace}))) \tag{9}$$

$$\begin{aligned} L_B = -\sum_{i=0}^{size(Y)} (\Delta Y_i log(\Delta P_i) \\ +(1-\Delta Y_i)log(1-\Delta P_i)) \end{aligned} \tag{10}$$

The Laplace operator is a second order differential operator in the n-dimensional Euclidean space, defined as the divergence of the gradient (Δf). Because the second derivative can be used to detect edges, we use the Laplace operator to get salient object boundaries. The Laplace operator in two dimensions is given by Eq.8, where x and y are the standard Cartesian coordinates of the xy-plane. In fact, since the Laplacian uses the gradient of images, it calls internally the convolution operation to perform its computation. Then we use absolute operation followed by tanh activatioin Eq.9 map the value to [0,1]. Finally we use the cross-entropy loss to supervise the generation of salient object boundaries Eq.10. The total loss function is their weighted sum:

$$L = \alpha L_S + (1-\alpha) L_B \tag{11}$$

4. Experiments

4.1. Datasets and Evaluation Criteria

The performance evaluation is utilized on five standard benchmark datasets: DUTS-test[30], ECSSD[37], HKU-IS[19], PASCAL-S[21] and DUT-OMRON[40]. DUTS[30] is a large scale dataset, which contains 10553 images for training and 5019 images for testing. ECSSD [37] contains 1,000 images with many semantically meaningful and complex structures in their ground truth segmentation. HKU-IS [19] contains 4447 challenging images with multiple disconnected salient objects, overlapping the image boundary or low color contrast. PASCAL-S [21] contains 850 images, different salient objects are labeled with different saliencies. DUT-OMRON [40] has 5,168 high quality images. Images of this dataset have one or more salient objects and relatively complex background.

Same as other state-of-the-art salient object detection methods, three popular criteria are used for performance evaluation, i.e. precision and recall curve (denoted PR curve), F-measure, weighted F-measure (denoted wF_β), and mean absolute error (MAE).

The precision and recall are computed by comparing the binary map under different thresholds between predicted saliency map and ground truth, the thresholds are from 0 to 255. wF_β is a overall evaluation standard computed by the weighted combination of precision and recall:

$$F_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{12}$$

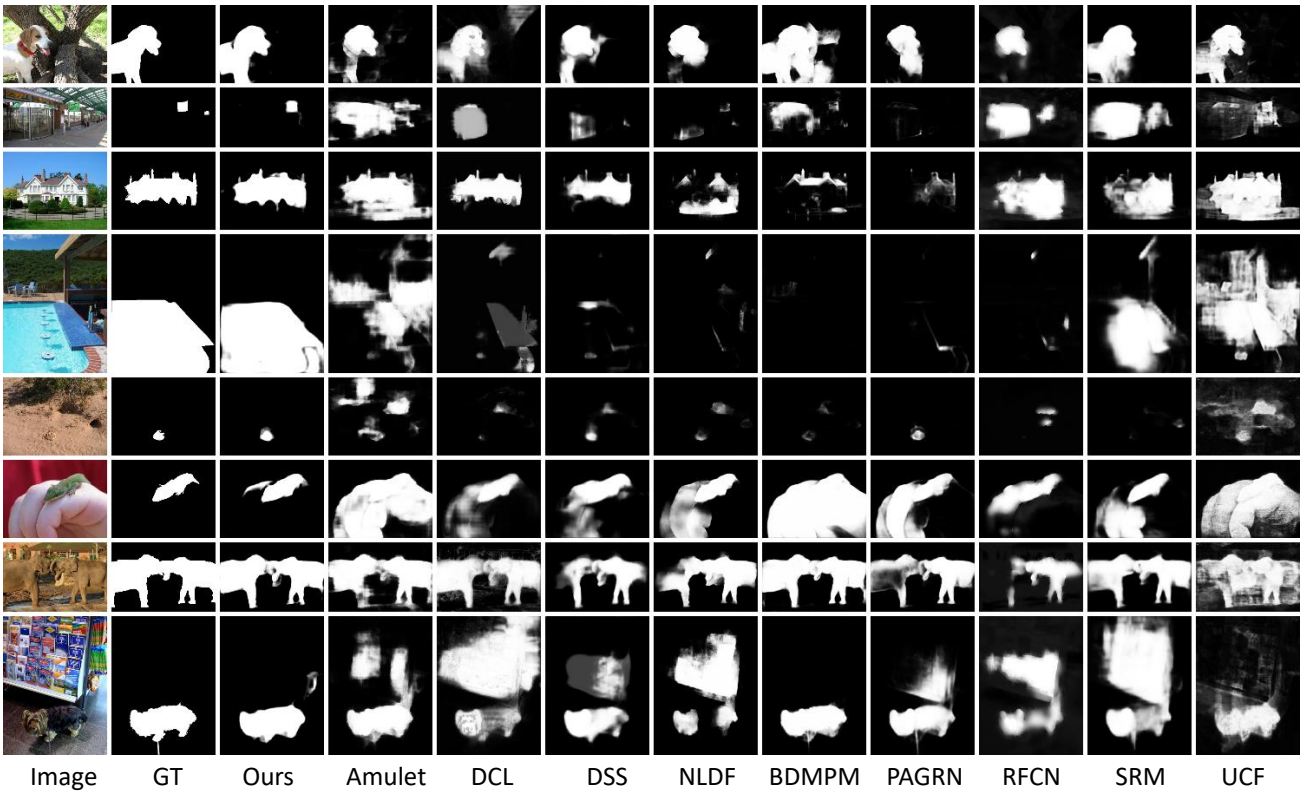

Figure 5. Visual comparisons of the proposed method and the state-of-the-art algorithms.

Where $\beta^2 = 0.3$ as used in other approaches. Mean absolute error (MAE) is computed by:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |P(x,y) - Y(x,y)| \qquad (13)$$

where Y is the ground truth(GT), and P is the saliency map of network output.

4.2. Implementation Details

We use VGG-16 pre-trained on Imagenet[7] as basic model. The DUTS-train dataset is used to train our model, which contains 10553 images. As suggested in [22], we don't use the validation set and train the model until training loss converges. To make the model robust, we adopt some data augmentation techniques: random rotating, random cropping, random brightness, saturation and contrast changing, and random horizontal flipping.

When training, we set α = 1.0 at beginning to generate rough saliency map. In this period, our model is trained using SGD[2] with an initial learning rate 1e-2, the image size is 256×256 , the batch size is 22. Then we adjust different α to refine the boundaries of saliency map,and find α = 0.7 is the optimal setting in experiment Tab.2. In this period, the image size, batch size is same as the previous period, but the initial learning rate is 1e-3. The code will be found at `https://github.com/CaitinZhao/cvpr2019_Pyramid-Feature-Attention-Network-for-Saliency-dete`

4.3. Comparison with State-of-the-arts

The performance of the proposed method is compared with eleven state-of-the-art salient object detection approaches on five test datasets, including BDMPM [42], GRL [33], PAGRN [45], Amulet [43], SRM [32], UCF [44], DCL [20], DHS [22], ELD [18], NLDF [24] and RFCN [31]. For fair comparisons, we use the implementations with recommended parameters and the saliency maps provided by the authors.

4.3.1 Visual Comparison

Fig.5 provides a visual comparison of our method and other state-of-the-arts. From Fig.5, our method gets the best detection results which are much close to the ground truth in various challenging scenarios. To be specific, (1) the proposed method not only highlights the correct salient object regions clearly, but also well suppresses the saliencies

Table 1. The wF_β and MAE of different salient object detection approaches on all test datasets. The best three results are shown in red, blue, and green.

Methods	DUTS-test		ECSSD		HKU-IS		PASCAL-S		DUT-OMRON	
	wF_β	MAE	wF_β	MAE	wF_β	MAE	wF_β	MAE	wF_β	MAE
Ours	0.8702	0.0405	0.9313	0.0328	0.9264	0.0324	0.8922	0.0677	0.8557	0.0414
BDMPM[42]	0.8508	0.0484	0.9249	0.0478	0.9200	0.0392	0.8806	0.0788	0.7740	0.0635
GRL[33]	0.8341	0.0509	0.9230	0.0446	0.9130	0.0377	0.8811	0.0799	0.7788	0.0632
PAGRN[45]	0.8546	0.0549	0.9237	0.0643	0.9170	0.0479	0.8690	0.0940	0.7709	0.0709
Amulet[43]	0.7773	0.0841	0.9138	0.0604	0.8968	0.0511	0.8619	0.0980	0.7428	0.0976
SRM[32]	0.8269	0.0583	0.9158	0.0564	0.9054	0.0461	0.8677	0.0859	0.7690	0.0694
UCF[44]	0.7723	0.1112	0.9018	0.0704	0.8872	0.0623	0.8492	0.1099	0.7296	0.1203
DCL[20]	0.7857	0.0812	0.8959	0.0798	0.8899	0.0639	0.8457	0.1115	0.7567	0.0863
DHS[22]	0.8114	0.0654	0.9046	0.0622	0.8901	0.0532	0.8456	0.0960	-	-
DSS[15]	0.8135	0.0646	0.8959	0.0647	0.9011	0.0476	0.8506	0.0998	0.7603	0.0751
ELD[18]	0.7372	0.0924	0.8674	0.0811	0.8409	0.0734	0.7882	0.1228	0.7195	0.0909
NLDF[24]	0.8125	0.0648	0.9032	0.0654	0.9015	0.0481	0.8518	0.1004	0.7532	0.0796
RFCN[31]	0.7826	0.0893	0.8969	0.0972	0.8869	0.0806	0.8554	0.1159	0.7381	0.0945

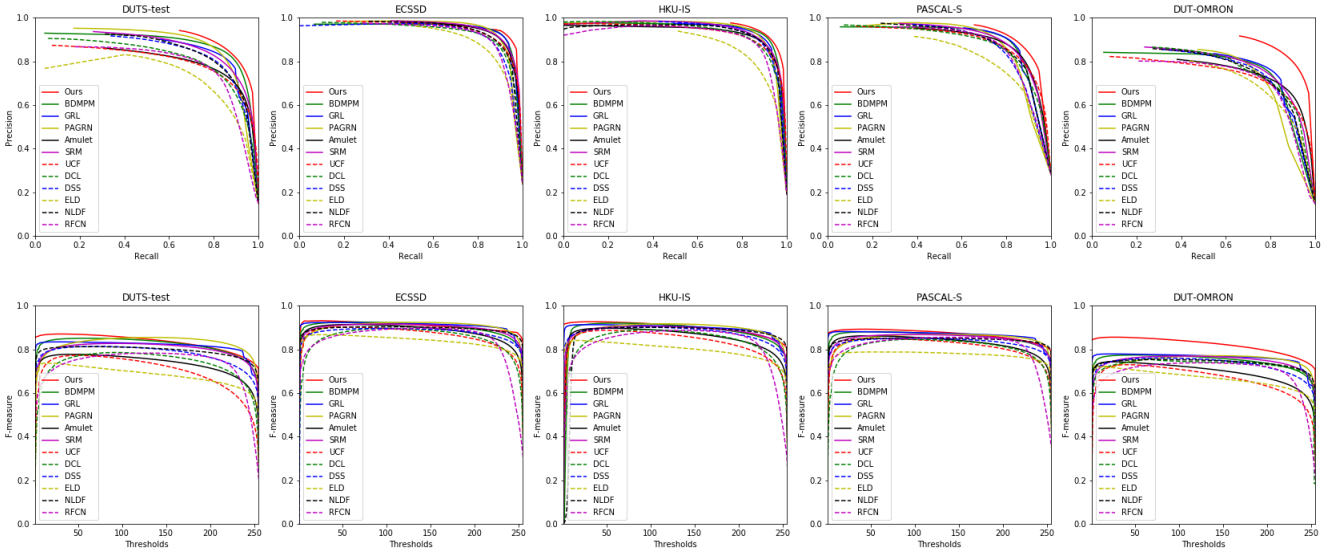

Figure 6. Quantitative comparisons of the proposed approach and eleven state-of-the-art CNN based salient object detection approaches on five datasets. The first and second rows are the PR curves and F-measure curves of different methods respectively.

of background regions, so as to produce the detection results with higher contrast between salient objects and background than other approaches. (2) With the help of the edge preservation loss, the proposed method is able to generate the salient maps with clear boundaries and consistent saliencies. (3) The saliency maps are much better than other works when salient objects are similar to background (Fig.5 the 2,5,7 rows) and the salient objects have special semantic information(Fig.5 the 1,3,4,6,8 rows).

4.3.2 Quantitative Comparison

Fig.6 and Tab.1 provides the quantitative evaluation results of the proposed method and eleven state-of-the-art salient object detection approaches on five test datasets in terms of PR curve, F-measure curve, wF_β and MAE criteria. As seen from Tab.1, our method gets the best result on five test datasets in terms of wF_β and MAE, which demonstrate the efficiency of the proposed method. From Fig.6, the PR curve and F-measure curve of our method are significantly higher than other methods, which means our method is more robust than other approaches even on challenging datasets. To be specific, our method gets larger improvement compared with the best existing approach on DUT-OMRON dataset. DUT-OMRON dataset is a difficult and challenging saliency detection dataset, in which there are many complex natural scenes images and the color of

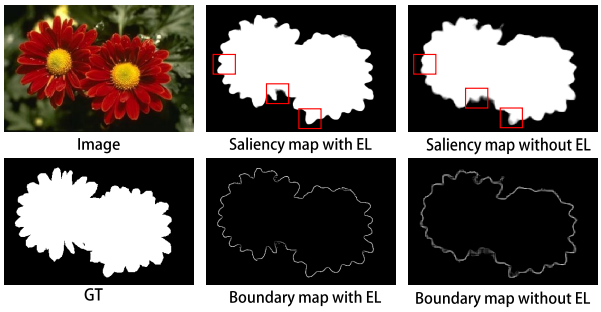

Figure 7. Visual comparison of saliency detection results with and without the edge preservation loss.

α	1.	0.9	0.8	0.7	0.6
wF_β	0.8528	0.8576	0.8602	0.8702	0.8619
MAE	0.0432	0.0427	0.0393	0.0405	0.0428

Table 2. The effectiveness of edge preservation loss. The score of wF_β and MAE in our method when α is given different values. The best result is shown in red. The test dataset is DUTS-test.

salient objects is similar to the background. The proposed method can effectively find correct salient objects with powerful feature extraction capability and apt attention mechanisms, which make the network focus on salient objects.

4.4. The Effectiveness of edge preservation loss

In Sec.3.3 we propose an effective edge preservation loss to guide network to learn more detailed information in boundary localization. Fig.7 shows the saliency maps generated from our method and boundary maps calculated by Eq.9 with edge preservation loss or not. These results illustrate that the edge preservation loss directly enhances the generality and make our method with fine details. In addition, we found that the edge preservation loss with different α have different effects on the final results. From Tab.2, when α is 0.7 gets the best result.

4.5. Ablation Study

To investigate the importance of different modules in our method, we conduct the ablation study. From Tab.3, that the proposed model contains all components (i.e. context-aware pyramid feature extraction(CPCE), channel-wise attention(CA), spacial attention(SA) and edge preservation loss(EL)) achieves the best performance, which demonstrates that all components are necessary for the proposed method to get the best salient object detection result.

We adopt the model only use high-level features as basic model, and the base MAE is 0.1003. First, we add CPFE to basic model and get decline in MAE, furthermore we add CA and get decline of 37% in MAE compared with basic model. Then we add low-level features to high-level features and prove the effectiveness of Integrating multi-scale features. On this basis, we add SA to low-level features and get decline of 57% in MAE compared with basic model. Finally, we add EL in the model and get the best result which get decline of 60% in MAE compared with basic model.

HL	CPFE	CA	LL	SA	EL	MAE
✓						0.1003
✓	✓					0.0815
✓	✓	✓				0.0629
✓			✓			0.0836
✓			✓		✓	0.0800
✓	✓	✓	✓			0.0528
✓	✓	✓	✓	✓		0.0432
✓	✓	✓	✓	✓	✓	**0.0405**

Table 3. Ablation Study using different components combinations. HL means use High-Level features, CPFE means use Context-aware pyramid Feature Extraction after high-level features, CA means use Channel-wise Attention after high-level features, LL means use Low-Level features, SA means use Spacial Attention after low-level features and EL means use Edge preservation Loss.

5. Conclusions

In this paper, we propose a novel salient object detection method named Pyramid Feature Attention network. In consideration of the different characteristics of different level features, for high-level features we design a context-aware pyramid feature extraction module contains different atrous convolutions at multi scales and a channel-wise attention module to capture semantic high-level features; For low-level features, we employ a spatial attention module to suppress the noises in background and focus on salient objects. Besides, we propose a novel edge preservation loss to guide network to learn more detailed information in boundary localization. In a word, the proposed method is expert in locating correct salient objects with powerful feature extraction capability and apt attention mechanisms, which make the network robost and effective in saliency detection. Experimental results on five datasets demonstrate that our proposed approach outperforms state-of-the-art methods under different evaluation metrics.

Acknowledgments: This work was supported in part by the Natural Science Foundation of China under Grant 61672194, by the National Key R&D Program of China under Grant 2018YFC0832304, by the Distinguished Youth Science Foundation of Heilongjiang Province of China under Grant JC2018021, and by the State Key Laboratory of Robotics and System (HIT) under Grant SKLRS-2019-KF-14.

References

[1] A. Borji, S. Frintrop, D. N. Sihite, and L. Itti. Adaptive object tracking by learning background context. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2012 IEEE Computer Society Conference on*, pages 23–30. IEEE, 2012.

[2] L. Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[3] L. Chen, H. Zhang, J. Xiao, L. Nie, J. Shao, W. Liu, and T.-S. Chua. Sca-cnn: Spatial and channel-wise attention in convolutional networks for image captioning. pages 6298–6306, 2017.

[4] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille. Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs. *IEEE transactions on pattern analysis and machine intelligence*, 40(4):834–848, 2018.

[5] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu. Global contrast based salient region detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 37(3):569–582, 2015.

[6] X. Chu, W. Yang, W. Ouyang, C. Ma, A. L. Yuille, and X. Wang. Multi-context attention for human pose estimation. *arXiv preprint arXiv:1702.07432*, 1(2), 2017.

[7] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Feifei. Imagenet: A large-scale hierarchical image database. pages 248–255, 2009.

[8] Y. Ding, J. Xiao, and J. Yu. Importance filtering for image retargeting. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 89–96. IEEE, 2011.

[9] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof. Saliency driven total variation segmentation. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 817–824. IEEE, 2009.

[10] Y. Gao, M. Wang, D. Tao, R. Ji, and Q. Dai. 3-d object retrieval and recognition with hypergraph analysis. *IEEE Transactions on Image Processing*, 21(9):4290–4303, 2012.

[11] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin. Convolutional sequence to sequence learning. *arXiv preprint arXiv:1705.03122*, 2017.

[12] D. Gilbarg and N. S. Trudinger. Elliptic partial differential equations of second order. 2001.

[13] J. He, J. Feng, X. Liu, T. Cheng, T.-H. Lin, H. Chung, and S.-F. Chang. Mobile product search with bag of hash bits and boundary reranking. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 3005–3012. IEEE, 2012.

[14] S. Hong, T. You, S. Kwak, and B. Han. Online tracking by learning discriminative saliency map with convolutional neural network. In *International Conference on Machine Learning*, pages 597–606, 2015.

[15] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr. Deeply supervised salient object detection with short connections. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5300–5309. IEEE, 2017.

[16] J. Hu, L. Shen, and G. Sun. Squeeze-and-excitation networks. *arXiv preprint arXiv:1709.01507*, 7, 2017.

[17] Z. Jiang and L. S. Davis. Submodular salient region detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2043–2050, 2013.

[18] G. Lee, Y.-W. Tai, and J. Kim. Deep saliency with encoded low level distance map and high level features. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 660–668, 2016.

[19] G. Li and Y. Yu. Visual saliency based on multiscale deep features. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5455–5463, 2015.

[20] G. Li and Y. Yu. Deep contrast learning for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 478–487, 2016.

[21] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille. The secrets of salient object segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 280–287, 2014.

[22] N. Liu and J. Han. Dhsnet: Deep hierarchical saliency network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 678–686, 2016.

[23] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[24] Z. Luo, A. K. Mishra, A. Achkar, J. A. Eichel, S. Li, and P.-M. Jodoin. Non-local deep features for salient object detection. In *CVPR*, volume 2, page 7, 2017.

[25] V. Mnih, N. Heess, A. Graves, et al. Recurrent models of visual attention. In *Advances in neural information processing systems*, pages 2204–2212, 2014.

[26] C. Peng, X. Zhang, G. Yu, G. Luo, and J. Sun. Large kernel matters improve semantic segmentation by global convolutional network. *computer vision and pattern recognition*, pages 1743–1751, 2017.

[27] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[28] Y. Tang and X. Wu. Saliency detection via combining region-level and pixel-level predictions with cnns. In *European Conference on Computer Vision*, pages 809–825. Springer, 2016.

[29] Y. Tang, X. Wu, and W. Bu. Deeply-supervised recurrent convolutional neural network for saliency detection. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 397–401. ACM, 2016.

[30] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan. Learning to detect salient objects with image-level supervision. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(CVPR)*, pages 136–145, 2017.

[31] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan. Saliency detection with recurrent fully convolutional networks. In *European Conference on Computer Vision*, pages 825–841. Springer, 2016.

[32] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu. A stage-wise refinement model for detecting salient objects in im-

ages. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4019–4028, 2017.

[33] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji. Detect globally, refine locally: A novel approach to saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3127–3135, 2018.

[34] H. Xu and K. Saenko. Ask, attend and answer: Exploring question-guided spatial attention for visual question answering. In *European Conference on Computer Vision*, pages 451–466. Springer, 2016.

[35] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[36] K. Xu, J. Ba, R. Kiros, K. Cho, A. Courville, R. Salakhudinov, R. Zemel, and Y. Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057, 2015.

[37] Q. Yan, L. Xu, J. Shi, and J. Jia. Hierarchical saliency detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1155–1162, 2013.

[38] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013.

[39] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013.

[40] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3166–3173, 2013.

[41] Z. Yang, X. He, J. Gao, L. Deng, and A. Smola. Stacked attention networks for image question answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 21–29, 2016.

[42] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang. A bi-directional message passing model for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1741–1750, 2018.

[43] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan. Amulet: Aggregating multi-level convolutional features for salient object detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 202–211, 2017.

[44] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin. Learning uncertain convolutional features for accurate saliency detection. In *Computer Vision (ICCV), 2017 IEEE International Conference on*, pages 212–221. IEEE, 2017.

[45] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang. Progressive attention guided recurrent network for salient object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 714–722, 2018.