electronics

MDPI

Article

Salient Object Detection Combining a Self-Attention Module and a Feature Pyramid Network

Guangyu Ren, Tianhong Dai, Panagiotis Barmpoutis and Tania Stathaki *

Department of Electrical and Electronic Engineering, Imperial College London, London SW7 2AZ, UK; g.ren19@imperial.ac.uk (G.R.); tianhong.dai15@imperial.ac.uk (T.D.); p.barmpoutis@imperial.ac.uk (P.B.)
* Correspondence: t.stathaki@imperial.ac.uk

Received: 4 September 2020; Accepted: 13 October 2020; Published: 16 October 2020

Abstract: Salient object detection has achieved great improvements by using the Fully Convolutional Networks (FCNs). However, the FCN-based U-shape architecture may cause dilution problems in the high-level semantic information during the up-sample operations in the top-down pathway. Thus, it can weaken the ability of salient object localization and produce degraded boundaries. To this end, in order to overcome this limitation, we propose a novel pyramid self-attention module (PSAM) and the adoption of an independent feature-complementing strategy. In PSAM, self-attention layers are equipped after multi-scale pyramid features to capture richer high-level features and bring larger receptive fields to the model. In addition, a channel-wise attention module is also employed to reduce the redundant features of the FPN and provide refined results. Experimental analysis demonstrates that the proposed PSAM effectively contributes to the whole model so that it outperforms state-of-the-art results over five challenging datasets. Finally, quantitative results show that PSAM generates accurate predictions and integral salient maps, which can provide further help to other computer vision tasks, such as object detection and semantic segmentation.

Keywords: salient object detection; pyramid self-attention module; fully convolution network; feature pyramid network

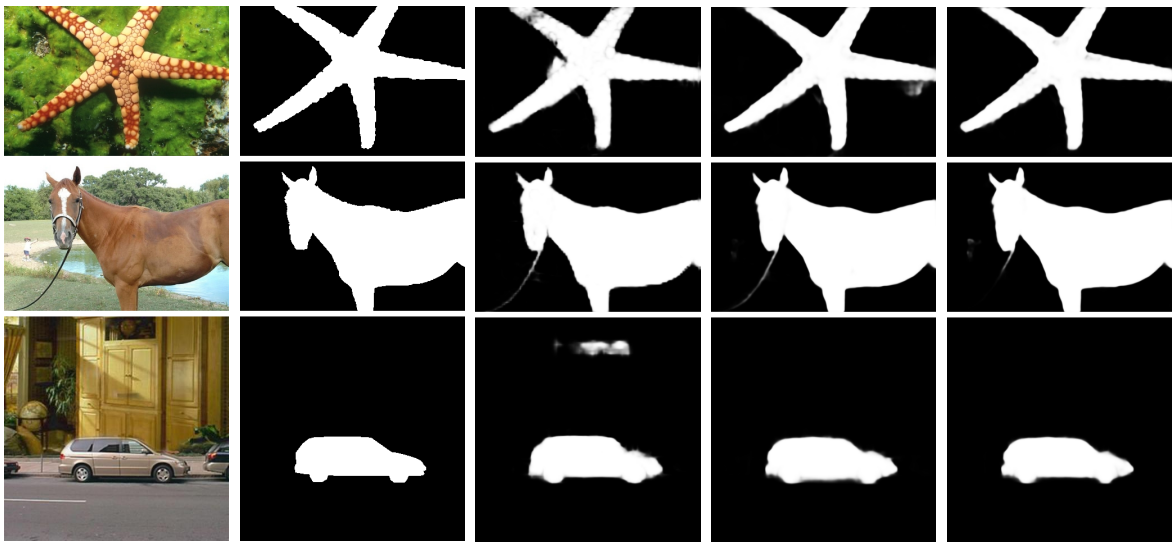
1. Introduction

Salient object detection or segmentation aims at identifying visually distinctive parts of a natural scene. With the capacity to provide high-level information, saliency detection problems are widely employed in various computer vision applications, such as object detection [1–3] and tracking [4], visual robotic manipulations [5,6], image segmentation [7,8] and video summarization [9,10]. In early studies, salient object detection was formulated as a binary segmentation problem. However, the connections between the salient object detection and other computer vision tasks was ambiguous. Nowadays, convolutional neural networks (CNNs) attract more attention in the research community. Compared with the classic hand-crafted feature descriptors [11,12], CNNs have robust feature representation ability. Specifically, CNN kernels with small receptive fields can provide local information, and the kernels with large receptive fields can provide global information. This characteristic enables CNN-based approaches to detect salient areas with refined boundaries [13]. Thus, CNN-based approaches have, in previous years, become the major research approach for salient object detection problems.

Furthermore, the so-called Fully Convolutional Networks (FCNs) have recently become the fundamental framework for the above-mentioned problems [14–16], as FCNs can achieve fewer parameters and smaller flexible input size compared to the fully connected layer. However, although these works have achieved great improvements in performance, they are still restricted by some limitations. FCN-based approaches utilize multiple convolution layers and pooling layers

to extract high-level semantic features. These features facilitate the location of objects, however, during pooling operations, important information might be lost. This can lead to degraded boundaries of detected objects being generated. Besides, when the high-level features are upsampled to generate the score prediction for each pixel, the features will also be diluted, which could decrease the ability of object localization.

In this paper, we propose a novel pyramid self-attention module (PSAM) to overcome the limitation of feature dilution in the previous FCN-based approaches. Figure 1c shows the inherent problems of Feature Pyramid Networks (FPNs) [17]. Through incorporating a self-attention module with multi-scale feature maps of FPNs, the model will focus on the high-level features. This leads to the extraction of features with richer high-level semantic information and larger receptive fields. In addition, a channel-wise attention module is employed to eliminate the redundancy in the FPN, which can refine the final results. Experimental results show that the proposed PSAM can improve the performance of salient object detection and achieve superior to state-of-the-art results in five challenging datasets. To summarise, the contributions of this work can be concluded as: (1) We propose a novel pyramid self-attention structure, which can make the model focus more on high-level features and reduce feature dilation in top-down pathways. (2) We adopt a channel-wise attention to reduce the redundant information in the lateral connections of the FPN to refine the final results. The code can be found in Supplementary Materials.



(**a**) Image (**b**) ground truth (**c**) baseline (**d**) baseline+PSAM (**e**) ours

Figure 1. Qualitative visual results of ablation studies. (**a**) original images, (**b**) ground truth, (**c**) baseline: a Feature Pyramid Network (FPN) with ResNet-50 as backbone (**d**) baseline+ Pyramid self-attention Module (PSAM): an FPN structure and a proposed pyramid self-attention module (**e**) ours: the final model which contains both a PSAM and channel-wise attention modules.

2. Related Works

2.1. Salient Object Detection

Different from the object detection, which is used to detect object instances with a specific category (e.g., car, bike, people.) [18], salient object detection is used to detect and segment the salient objects in natural images [13]. Due to the outstanding feature representation ability of CNNs, the handcraft feature-based methods have been replaced by the CNN models. In the work of [19], Li and Yu used fully connected layers on the top CNN layers to extract different scale features of a single region. Then, multi-scale features were used to predict the scores for each region. Reference [20] utilized two independent CNNs to extract both the global context of the entire image and the detailed context of

the local area under consideration to train the model jointly. However, the spatial information was lost in these CNN-based methods, because of the fully connected layers.

Recently, FCN-based methods have raised more interest in the salient object detection. Reference [15] proposed a boundary-aware salient object detection network which incorporates a prediction module and a residual refinement module (RRM). The prediction module was used to estimate the salient map from the raw images and the RRM was used to refine the results from the prediction module which was trained between the salient map and the ground truth. Reference [14] introduced a so-called PoolNet structure with two pooling modules: a global guidance module (GGM) and a feature aggregation module (FAM). The GGM was designed to acquire more high-level information around the inputs, which tackles the feature dilation problem in the U-shape network structure. Furthermore, the FAM merges the multi-scale features of the FPN, leading it in reducing the problem of aliasing caused by the up-sampling and enlarging the receptive fields. From the experiments, it is demonstrated that PoolNet can make a more accurate localization of the sharpened salient objects compared to other baseline approaches. In the work of [21], a Cascaded Partial Decoder (CPD) structure that contains two prime branches was proposed. The first branch contributes to the computation speed improvement by dropping features in the shallow layers. The second branch uses the salient map from the first branch to refine the features in the deeper layers, which ensures the speed and accuracy of the framework.

2.2. Attention Mechanism

Attention mechanism, an approach to solve the problem of the loss of relevant information, is extensively used in the area of Natural Language Processing (NLP). Reference [22] introduced a framework called a transformer, a new encoder–decoder architecture that uses only the attention mechanism instead of recurrent layers to encode each position to capture global dependencies between input and output. This framework also allows parallel computing, which leads to a faster computational speed compared to recurrent networks. Except for the sequence models, this kind of attention mechanism is also needed in CNN models. Different from the attention mechanism in the sequence models, the self-attention mechanism pays more attention to a single datum. Reference [23] proposed the Attention Augmented Convolutional Network, which produces attentional feature maps via a self-attention module and concatenates these with CNN feature maps to capture the spatial dependencies of the input, and it achieves a huge improvement in the tasks of object classification and detection. The stand-alone self-attention layer was introduced in the work of [24]. It can be used to set up a fully self-attentional model through replacing all spatial convolution layers with self-attention layers. The previous works proves that it can be used as an effective stand-alone layer which can replace the spatial convolution layer easily. Reference [25] proposed a residual attention network combined with multiple attention modules to deal with the classification task. Instead of stacking attention modules directly in the network, a residual attention mechanism is employed to train the neural network with hundreds of layers. With an increasing number of layers, the attention features adjust adaptively and the proposed network achieves a better performance and lower computational complexity.

3. The Proposed Method

In this section, we describe the proposed novel architecture that integrates the two attention modules. More specifically, as shown in Figure 2, we use a pyramid self-attention module which aims to enhance the high-level semantic features and transmit the enhanced semantic information to different feature levels. In addition, when feature maps are merged in the top-down pathway, a simple channel-wise attention [26–28] module is added in each lateral connection to focus on the salient objects in feature maps.

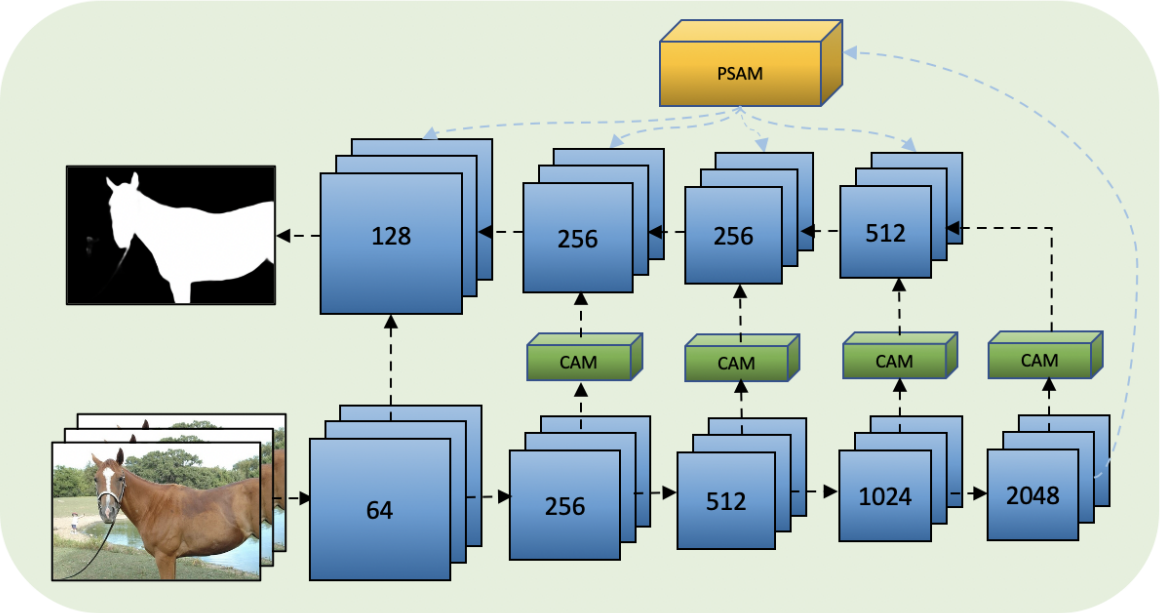


Figure 2. Overall pipeline of the proposed model: a pyramid self-attention module (PSAM) is built on the basic FPN structure and four channel-wise attention (CAM) modules are added in lateral connections.

3.1. Pyramid Self-Attention Module

In this subsection, we describe the proposed module framework in detail and demonstrate the differences from previous works. Reference [29] has demonstrated that high-level semantic features are more representative and discriminative, leading to the position of salient objects being located more accurately. Figure 1c shows that, without any additional attention modules, the FPN baseline can generate rough saliency maps which have incomplete salient objects. Meanwhile, there are also some non-salient objects present, which should not be detected in the saliency maps. These error predictions are caused by two main challenges which cannot be avoided in the FPN architecture. The first problem is that the high-level information is diluted progressively when it is integrated in different feature levels in the top-down pathway. Another problem of the FPN baseline is that this architecture can be impacted by other non-essential information, which may reduce the final performance of the model. In other words, the FPN architecture detects not only incomplete salient objects but also irrelevant objects. To overcome these two intrinsic problems of the baseline, we propose a novel pyramid self-attention module (PSAM) which contains stand-alone self-attention layers [24] in different scales, further focusing on important regions and enlarging the receptive field of the model. Specifically, as shown in Figure 3, PSAM firstly transforms the feature map which is produced by a bottom-up pathway into multi-scale feature regions, and then each self-attention layer learns to pay more attention on important semantic information. After being processed by self-attention layers, these multi-scale representations, which contain effective semantic information, are concatenated together to complement high-level semantic information in the top-down pathway. More technically, let X^{in} denote the feature map, which is produced by the top-most layer. We downsample the feature map $X^{in} \in R^{H\times\ W\times C}$ into three different scales, denoted as $\{X_1, X_2, X_3\}$, where H, W and C represent the height, width and channels of the feature map, respectively. Given a pixel $x_{ij} \in \{X_1, X_2, X_3\}$, a corresponding local memory block r_k^{ij} is extracted from the same feature map $X_i \in \{X_1, X_2, X_3\}$. This r_k^{ij} is a $k \times k$ region which surrounds x_{ij}. There are three crucial learnable parameters in this

self-attention algorithm, namely, queries, keys and values. We use W_Q, W_K and W_V to represent their learnable weights respectively. The final attention output pixel is computed as follows

$$s_{ij} = \sum_{a,b \in N_k} \text{softmax}(q_{ij}^T k_{ab}) v_{ab}, \tag{1}$$

where $q_{ij} = W_Q x_{ij}$, $k_{ab} = W_K x_{ab}$, $v_{ab} = W_V x_{ab}$ denote the three crucial parameters, s_{ij} denotes the output pixel of a self-attention layer, $N_k = \{(a,b) : a = \{-\frac{k-1}{2}, ..., \frac{k-1}{2}\}, b = \{-\frac{k-1}{2}, ..., \frac{k-1}{2}\}\}$ defines the coordinates x_{ab} of r_k^{ij} and we use $\{Y_1, Y_1, Y_3\}$ to denote the final feature map, which further represents an upsampling operation after each self-attention layer. Then, we concatenate them with the original X^{in} to generate the final output of the PSAM.

$$Y^{out} = \text{concat}\{Y_1, Y_1, Y_3, X^{in}\} \tag{2}$$

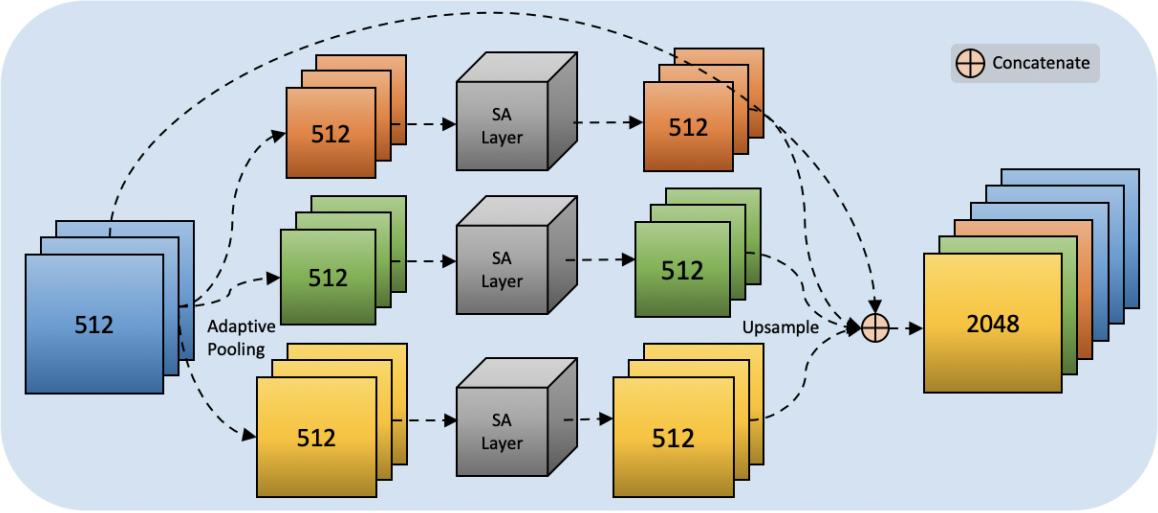


Figure 3. The structure of Pyramid Self-attention Module: red, green and yellow denote 1×1, 3×3 and 5×5 size, respectively. All feature maps in different size have 512 channels and finally concatenate together.

3.2. Channel-Wise Attention

To enhance the contextual and structural information, the lateral connection has been used in the top-down pathway, leading to a state-of-the-art performance of detection tasks. However, this operation also introduces some non-meaningful information, which can reduce the performance and impact the final prediction. From Figure 1c, it is observed that feature redundancy could cause two problems. The first problem is that there are extra regions which should not be detected in the saliency map. Another problem is that the edges of salient objects are ambiguous. Both problems indicate that a further refinement should be applied. Reference [27] has pointed out that different channels have different semantic features and channel-wise attention can capture channel-wise dependencies. In other words, channel-wise attention can emphasize the salient objects and alleviate the inaccuracy which is caused by redundant features in channels. Therefore, we add this simple channel-wise attention [26–28] to each later connection to achieve a refinement task. The structure of channel-wise attention is shown in Figure 4. It consists of one pooling layer and two fully connected layers, which are followed by

a ReLU [30] and a sigmoid function, respectively. First, an operation of squeezing global spatial information is applied to each channel. This step can be easily implemented by an average pooling

$$p_c = F_{pool}(X_c) = \frac{1}{H \times W} \sum_{i=1}^{H} \sum_{j=1}^{W} x_c(i,j) \tag{3}$$

where c refers to the channel number and $H \times W$ refers to the spatial dimensions of the i-th element of X_c. After the pooling operation, the generated channel descriptor is fed into the fully connected layers to fully capture channel-wise dependencies

$$s_c = F_{fc}(p_c, W) = \sigma(fc_2(\delta(fc_1(p_c, W_1)), W_2)) \tag{4}$$

where σ refers to the sigmoid function, δ refers to the Rectified Linear Unit (ReLU) function and fc denotes the fully connected layers. Finally, this generated scalar s_c multiplies the feature map X_c to generate a weighted feature map $\hat{X}_c$:

$$\hat{X}_c = s_c \cdot X_c \tag{5}$$

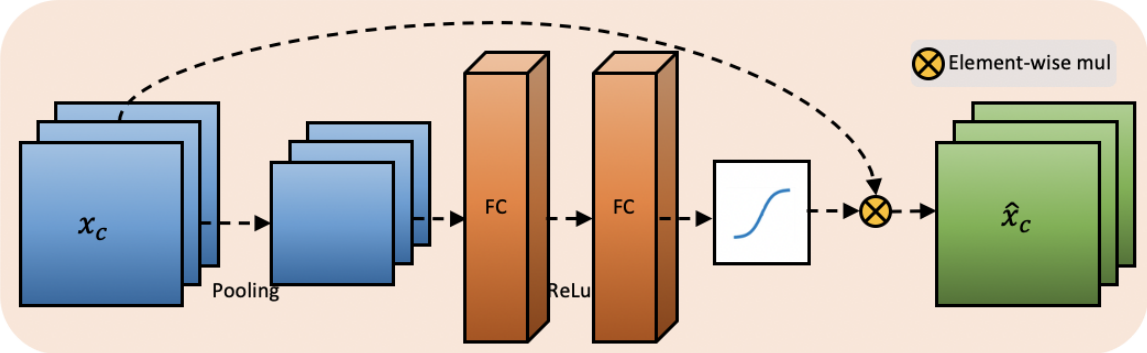


Figure 4. The structure of Channel-wise attention Module: feature maps experience an average pooling operation and squeeze to 128 channels to pass the first fully connected layer and retrieve to original channels in the second fully connected layer.

4. Experiment

4.1. Datasets and Evaluation Metrics

For the evaluation of the proposed methodology, we carry out a series of experiments using five popular saliency detection benchmarks, namely, ECSSD [31], DUT-OMRON [32], DUT-TE [33], HKU-IS [19] and SOD [34]. These five datasets consist of a variety of objects and structures which are challenging for salient object detection algorithms to locate and detect precisely. For the training of our model, we use the large-scale dataset DUTS [33], which contains 10,533 training images and 5019 testing images. To evaluate the performance of the model, we estimate three representative evaluation metrics, namely, precision–recall (PR) curves, F-score and mean absolute error (MAE). The PR curve is calculated by saliency prediction and ground truth:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \quad \text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{6}$$

where TP, FP and FN represent true-positive, false-positive and false-negative rates, respectively. Thresholds are set from 0 to 255 to binarize the prediction. These thresholds produce corresponding

precision–recall values to form a PR curve to represent the overall performance of the model. The *F*-score indicates the standard overall performance, which is computed by precision and recall

$$F_\beta = \frac{\left(1+\beta^2\right) \cdot \text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}} \tag{7}$$

where β^2 is set to 0.3 as default. Precision and recall are obtained by using different thresholds to compare prediction and ground truth. The MAE indicates the deviations between the binary saliency map and the ground truth. In other words, this metric quantifies the similarity between the prediction map and the ground truth mask as follows

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} |P(x,y) - G(x,y)| \tag{8}$$

where W denotes the width and H denotes the height of prediction, P denotes the prediction map, which is the output of the model, and G represents the ground truth map.

4.2. Impelmentation Details

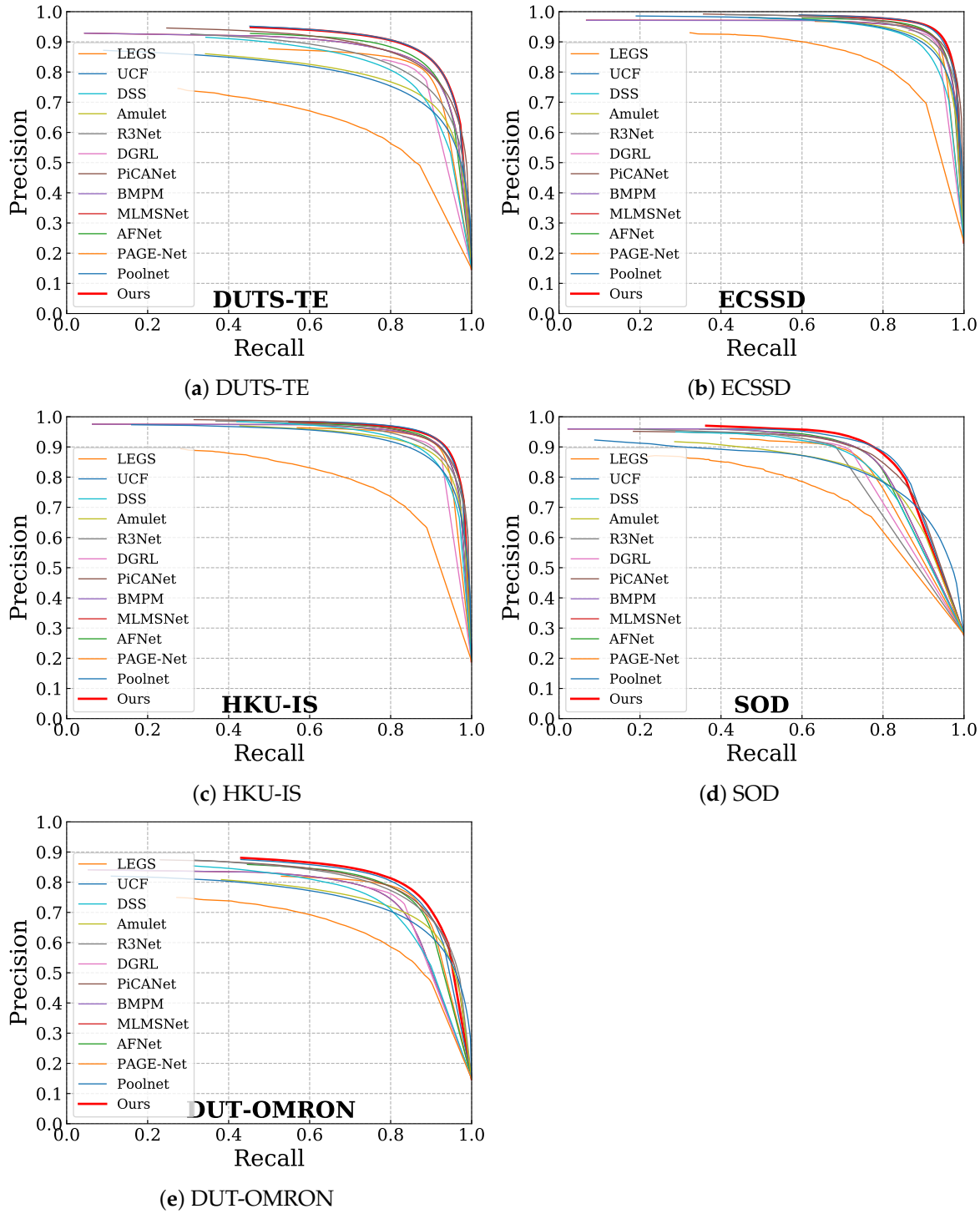
Our model is implemented in Pytorch. We use ResNet-50 [35] as a backbone which has been pre-trained on an ImageNet [36]. Specifically, the final stage exploits atrous convolution with dilated ratio 2 in order to obtain larger receptive fields. ImageNet contains 1.28 M images for training and 50 K for validation from 1000 classes. It is believed that pre-training on ImageNet could set reasonable initial parameters and accelerate the convergence of training in other tasks. The proposed architecture is trained on a GTX TITAN X GPU for 24 epochs. As suggested in [14], the initial learning rate is set equal to 5×10^{-5} for the first 15 epochs, and then reduces to 5×10^{-6} for the last nine epochs. We adopt a 0.0005 weight decay for the Adam [37] optimizer and binary cross entropy loss function in the proposed framework. Finally, to increase the robustness of the model, we perform data augmentation through the application of random horizontal flipping.

4.3. Comparisons with State-of-the-Arts

We perform our proposed method on five datasets to compare with 11 previous state-of-the-art methods, which include LEGS [38], UCF [39], DSS [29], Amulet [40], R3Net [41], DGRL [42], PiCANet [43], BMPM [44], MLMSNet [45], AFNet [46], PAGE-Net [16] and PoolNet [14]. For fair comparisons, we use the results which are generated by their original work with default parameters and released by the authors. Furthermore, all results are evaluated by the same evaluation method without any other processing tools.

4.3.1. Quantitative Comparisons

Figure 5 and Table 1 show the evaluation results of the proposed framework in comparison with eleven state-of-the-art methods on five challenging salient object datasets, as already mentioned above. More specifically, in Figure 5, it is clearly demonstrated from the PR curve that the proposed methodology represented by the red line outperforms the state-of-the-art methods. This result means that our method has better robustness than other previous methods. Furthermore, the quantitative results are listed in Table 1. The proposed method achieves a higher F-score and lower error scores than other methods, demonstrating that our novel model outperforms almost all previous state-of-the-art models on the different testing datasets. However, although it is noted that PoolNet has similar state-of-the-art results, our model has 23% less parameters.



(**a**) DUTS-TE (**b**) ECSSD

(**c**) HKU-IS (**d**) SOD

(**e**) DUT-OMRON

Figure 5. Results with PR curves on five benchmark datasets: DUTS-TE, ECSSD, HKU-IS, SOD and DUT-OMRON. x-axis represents the recall rate and y-axis represents the precision.

4.3.2. Qualitative Comparisons

Figure 6 illustrates the visual comparisons to further show the advantages of the proposed method. More precisely, compared to other approaches, the proposed method demonstrates the best performance on different challenging scenarios, with the results being closer to the corresponding ground truths compared to previous methods. For instance, in the first row, DGRL [42] and PiCANet [43] only detect the main body of the bird. LEGS [38], UCF [39], DSS [29] and Amulet [40]

detect irrelevant parts. Although R3Net [41] achieve a similar prediction to ours, the tail part still has low scores.

Table 1. Quantitative results with F-score and MAE on five challenging datasets: DUTS-TE, ECSSD, HKU-IS, SOD and DUT-OMRON. Our method is compared with 11 competitive baseline methods.

Method	DUTS-TE		ECSSD		HKU-IS		SOD		DUT-OMRON	
	F-Score	MAE	F-Score	MAE	F-Score	MAE	F-Score	MAE	F-Score	MAE
LEGS [38]	0.654	0.138	0.827	0.118	0.770	0.118	0.733	0.196	0.669	0.133
UCF [39]	0.771	0.117	0.910	0.078	0.888	0.074	0.803	0.164	0.734	0.132
DSS [29]	0.813	0.064	0.907	0.062	0.900	0.050	0.837	0.126	0.760	0.074
Amulet [40]	0.778	0.085	0.914	0.059	0.897	0.051	0.806	0.141	0.743	0.098
R3Net [41]	0.824	0.066	0.924	0.056	0.910	0.047	0.840	0.136	0.788	0.071
PiCANet [43]	0.851	0.054	0.931	0.046	0.922	0.042	0.853	0.102	0.794	0.068
DGRL [42]	0.828	0.050	0.922	0.041	0.910	0.036	0.845	0.104	0.774	0.062
BMPM [44]	0.851	0.048	0.928	0.045	0.920	0.039	0.855	0.107	0.774	0.064
PAGE-Net [16]	0.838	0.051	0.931	0.042	0.920	0.036	0.841	0.111	0.791	0.062
MLMSNet [45]	0.852	0.048	0.928	0.045	0.920	0.039	0.855	0.107	0.774	0.064
AFNet [46]	0.863	0.045	0.935	0.042	0.925	0.036	0.856	0.109	0.797	0.057
PoolNet [45]	0.880	0.040	0.944	0.039	0.933	0.032	0.870	0.101	0.803	0.056
Ours	0.879	0.040	0.944	0.038	0.931	0.034	0.874	0.104	0.813	0.056

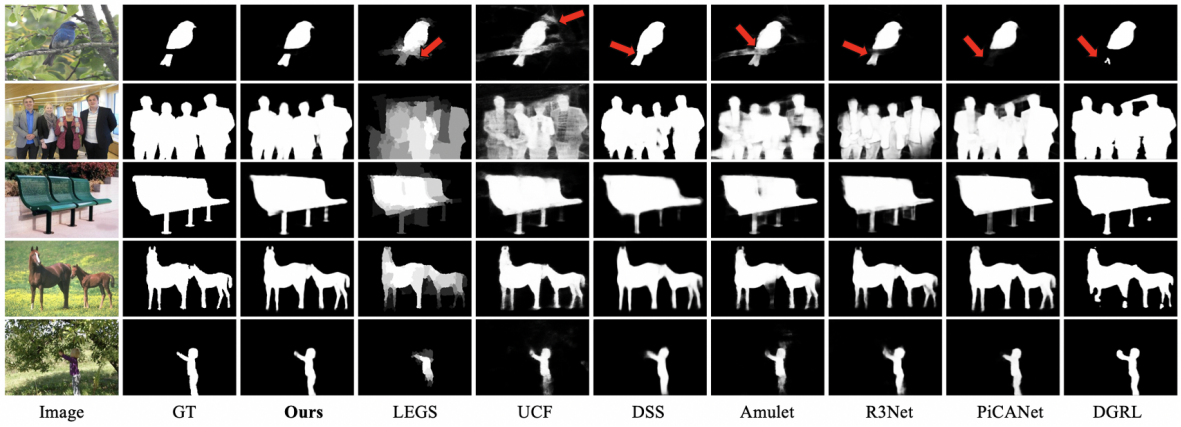


Figure 6. Overall comparison of qualitative visual results between our method and selected baseline methods. It shows that our method is able to provide a more complete salient map and smooth boundaries.

4.4. Ablation Study

In this subsection, we conduct a series of experiments on five different datasets to investigate the effectiveness of two modules. The ablation experiments are trained on the DUTS [33] training dataset in the same environment. From Table 2, the model which contains PSAM and channel-wise attention module achieves the best performance, demonstrating that the proposed modules can effectively assist the baseline's salient object detection performance. More specifically, we initially conduct the baseline experiments on the FPN baseline with ResNet-50 as a backbone. This basic model can generate rough saliency maps, such as, for example, the ones depicted in Figure 1c. Furthermore, we add a pyramid self-attention module (PSAM) on the baseline and the F-score increases significantly on all benchmark datasets, an observation which is more prominent for DUTS-TE [33] and SOD [34]. On this basis, we add channel-wise attention on the model to compose the proposed framework. The final results show that the channel-wise attention modules can further increase the performance and alleviate error predictions. To this end, Figure 1d,e demonstrate the effectiveness of two modules, respectively.

Table 2. Quantitative results with F-score and MAE on five challenging datasets: DUTS-TE, ECSSD, HKU-IS, SOD and DUT-OMRON for the ablation studies.

Method	DUTS-TE		ECSSD		HKU-IS		SOD		DUT-OMRON	
	F-Score	MAE	F-Score	MAE	F-Score	MAE	F-Score	MAE	F-Score	MAE
Baseline	0.856	0.045	0.933	0.045	0.921	0.037	0.848	0.116	0.785	0.059
Baseline+PSAM	0.876	0.041	0.940	0.042	0.928	0.034	0.857	0.121	0.803	0.056
Baseline+PSAM+CA	0.879	0.040	0.944	0.038	0.931	0.034	0.874	0.104	0.813	0.056

5. Discussion and Conclusions

As shown in Figure 1d,e and Table 2, it is observed that the proposed PSAM could enhance the semantic features and help the model focus on the useful objects. We believe that this attention module could provide further help to other computer vision tasks, such as object detection and segmentation, especially for semantic segmentation, which could facilitate multi-object saliency detection.

In this paper, we propose a novel end-to-end salient object detection method. Considering the intrinsic problems of the FPN architecture, a pyramid self-attention module (PSAM) is designed. This module contains different self-attention layers at multiple scales, leading in efficiently capturing multi-scale, high-level features, producing a model which focuses on the high-level semantic information and further enlarges the receptive field. Furthermore, we employ the channel-wise attention in lateral connections to reduce the feature redundancy and refine prediction results. Experimental results on five challenging datasets demonstrate that our proposed model surpasses 11 state-of-the-art methods. The ablation experiments further demonstrate the effectiveness of the two proposed modules.

Supplementary Materials: The code can be found in https://github.com/ic-qialanqian/PSAMNet.

Author Contributions: Conceptualization, methodology and writing—original draft preparation, G.R.; Project administration, writing—review editing and supervision, P.B. and T.S.; Methodology and software, T.D. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the EU H2020 TERPSICHORE project "Transforming Intangible Folkloric Performing Arts into Tangible Choreographic Digital Objects" under the grant agreement 691218.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Liu, T.; Huang, J.-J.; Dai, T.; Ren, G.; Stathaki, T. Gated multi-layer convolutional feature extraction network for robust pedestrian detection. *arXiv* **2019**, arXiv:1910.11761.
2. Ren, Z.; Gao, S.; Chia, L.-T.; Tsang, I.W.-H. Region-based saliency detection and its application in object recognition. *IEEE Trans. Circuits Syst. Video Technol.* **2013**, *24*, 769–779. [CrossRef]
3. Zhang, D.; Meng, D.; Zhao, L.; Han, J. Bridging saliency detection to weakly supervised object detection based on self-paced curriculum learning. *arXiv* **2017**, arXiv:1703.01290.
4. Hong, S.; You, T.; Kwak, S.; Han, B. Online tracking by learning discriminative saliency map with convolutional neural network. In Proceedings of the International Conference on Machine Learning, Lille, France, 6–11 July 2015; pp. 597–606.
5. Schillaci, G.; Bodiroža, S.; Hafner, V.V. Evaluating the effect of saliency detection and attention manipulation in human-robot interaction. *Int. J. Soc. Robot.* **2013**, *5*, 139–152. [CrossRef]
6. Yuan, X.; Yue, J.; Zhang, Y. Rgb-d saliency detection: Dataset and algorithm for robot vision. In Proceedings of the International Conference on Robotics and Biomimetics, Kuala Lumpur, Malaysia, 12–15 December 2018; pp. 1028–1033.
7. Wang, X.; You, S.; Li, X.; Ma, H. Weakly-supervised semantic segmentation by iteratively mining common object features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1354–1362.

8. Wei, Y.; Feng, J.; Liang, X.; Cheng, M.-M.; Zhao, Y.; Yan, S. Object region mining with adversarial erasing: A simple classification to semantic segmentation approach. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1568–1576.
9. Ma, Y.-F.; Lu, L.; Zhang, H.-J.; Li, M. A user attention model for video summarization. In Proceedings of the International Conference on Multimedia, 2002; pp. 533–542. Available online: https://dl.acm.org/doi/abs/10.1145/641007.641116 (accessed on 1 September 2020).
10. Simakov, D.; Caspi, Y.; Shechtman, E.; Irani, M. Summarizing visual data using bidirectional similarity. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Anchorage, AK, USA, 23–28 June 2008; pp. 1–8.
11. Dalal, N.; Triggs, B. Histograms of oriented gradients for human detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–25 June 2005; Volume 1, pp. 886–893.
12. Lowe, D.G. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* **2004**, *60*, 91–110. [CrossRef]
13. Borji, A.; Cheng, M.-M.; Hou, Q.; Jiang, H.; Li, J. Salient object detection: A survey. *Comput. Vis. Media* **2014**, *5*, 117–150. [CrossRef]
14. Liu, J.-J.; Hou, Q.; Cheng, M.-M.; Feng, J.; Jiang, J. A simple pooling-based design for real-time salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 3917–3926.
15. Qin, X.; Zhang, Z.; Huang, C.; Gao, C.; Dehghan, M.; Jagersand, M. Basnet: Boundary-aware salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 7479–7489.
16. 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, Long Beach, USA, 15–21 June 2019; pp. 1448–1457.
17. Lin, T.-Y.; Dollár, P.; Girshick, R.; He, K.; Hariharan, B.; Belongie, S. Feature pyramid networks for object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 2117–2125.
18. Zou, Z.; Shi, Z.; Guo, Y.; Ye, J. Object detection in 20 years: A survey. *arXiv* **2019**, arXiv:1905.05055.
19. Li, G.; Yu, Y. Visual saliency based on multiscale deep features. In Proceedings of the IEEE Conference on Computer vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 5455–5463.
20. 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, Boston, MA, USA, 7–12 June 2015; pp. 1265–1274.
21. Wu, Z.; Su, L.; Huang, Q. Cascaded partial decoder for fast and accurate salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–21 June 2019; pp. 3907–3916.
22. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems*; Morgan Kaufmann Pub: Burlington, MA, USA, 2017; pp. 5998–6008.
23. Bello, I.; Zoph, B.; Vaswani, A.; Shlens, J.; Le, Q.V. Attention augmented convolutional networks. In Proceedings of the IEEE International Conference on Computer Vision, Seoul, Korea, 27–28 October 2019; pp. 3286–3295.
24. Ramachandran, P.; Parmar, N.; Vaswani, A.; Bello, I.; Levskaya, A.; Shlens, J. Stand-alone self-attention in vision models. *arXiv* **2019**, arXiv:1906.05909.
25. Wang, F.; Jiang, M.; Qian, C.; Yang, S.; Li, C.; Zhang, H.; Wang, X.; Tang, X. Residual attention network for image classification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 3156–3164.
26. Chen, L.; Zhang, H.; Xiao, J.; Nie, L.; Shao, J.; Liu, W.; Chua, T.-S. 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, Honolulu, HI, USA, 21–26 July 2017; pp. 5659–5667.

27. Hu, J.; Shen, L.; Sun, G. Squeeze-and-excitation networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2017; pp. 7132–7141. Available online: https://arxiv.org/abs/1709.01507 (accessed on 1 September 2020).
28. Zhao, T.; Wu, X. Pyramid feature attention network for saliency detection. *arXiv* **2019**, arXiv:1903.00179.
29. Hou, Q.; Cheng, M.-M.; Hu, X.; Borji, A.; Tu, Z.; Torr, P.H. Deeply supervised salient object detection with short connections. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *41*, 815–828. [CrossRef] [PubMed]
30. Nair, V.; Hinton, G.E. Rectified linear units improve restricted boltzmann machines. In Proceedings of the International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010; pp. 807–814.
31. Yan, Q.; Xu, L.; Shi, J.; Jia, J. Hierarchical saliency detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 1155–1162.
32. 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, Portland, OR, USA, 23–28 June 2013; pp. 3166–3173.
33. Wang, L.; Lu, H.; Wang, Y.; Feng, M.; Wang, D.; Yin, B.; Ruan, X. Learning to detect salient objects with image-level supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 136–145.
34. Movahedi, V.; Elder, J.H. Design and perceptual validation of performance measures for salient object segmentation. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA 13–18 June 2010; pp. 49–56.
35. He, K.; Zhang, X.; Ren, S.; Sun, J. Deep residual learning for image recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 21–27 June 2016; pp. 770–778.
36. Krizhevsky, A.; Sutskever, I.; Hinton, G.E. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*; Morgan Kaufmann Pub: Burlington, MA, USA, 2012; pp. 1097–1105.
37. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. *arXiv* **2014**, arXiv:1412.6980.
38. Wang, L.; Lu, H.; Ruan, X.; Yang, M.-H. Deep networks for saliency detection via local estimation and global search. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 3183–3192.
39. 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, Venice, Italy, 22–29 October 2017; pp. 212–221.
40. 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, Venice, Italy, 22–29 October 2017; pp. 202–211.
41. Deng, Z.; Hu, X.; Zhu, L.; Xu, X.; Qin, J.; Han, G.; Heng, P.-A. R3net: Recurrent residual refinement network for saliency detection. In Proceedings of the International Joint Conference on Artificial Intelligence, Stockholm, Sweden, 13–19 July 2018; pp. 684–690.
42. Wang, T.; Zhang, L.; Wang, S.; Lu, H.; Yang, G.; Ruan, X.; Borji, A. Detect globally, refine locally: A novel approach to saliency detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Salt Lake City, UT, USA, 18–22 June 2018; pp. 3127–3135.
43. 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, Salt Lake City, UT, USA, 18–23 June 2018; pp. 3089–3098.
44. 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, Salt Lake City, UT, USA, 18–22 June 2018; pp. 1741–1750.
45. Wu, R.; Feng, M.; Guan, W.; Wang, D.; Lu, H.; Ding, E. A mutual learning method for salient object detection with intertwined multi-supervision. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 15–21 January 2019; pp. 8150–8159.

46. Feng, M.; Lu, H.; Ding, E. Attentive feedback network for boundary-aware salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Long Beach, USA, 15–21 June 2019; pp. 1623–1632.