Perception-and-Regulation Network for Salient Object Detection

Jinchao Zhu, Xiaoyu Zhang, Xian Fang, Yuxuan Wang, Panlong Tan, and Junnan Liu

Abstract—Effective fusion of different types of features is the key to salient object detection (SOD). The majority of the existing network structure designs are based on the subjective experience of scholars, and the process of feature fusion does not consider the relationship between the fused features and the highest-level features. In this paper, we focus on the feature relationship and propose a novel global attention unit, which we term the “perception-and-regulation” (PR) block, that adaptively regulators the feature fusion process by explicitly modelling the interdependencies between features. The perception part uses the structure of the fully connected layers in the classification networks to learn the size and shape of the objects. The regulation part selectively strengthens and weakens the features to be fused. An imitating eye observation module (IEO) is further employed to improve the global perception capabilities of the network. The imitation of foveal vision and peripheral vision enables the IEO to scrutinize highly detailed objects and organize a broad spatial scene to better segment objects. Sufficient experiments conducted on the SOD datasets demonstrate that the proposed method performs favourably against the 29 state-of-the-art methods.

Index Terms—Salient object detection, convolutional neural networks, attention mechanism, global perception.

I. INTRODUCTION

SALIENT object detection [1], [2], [3], [4], [5], [6] aims to find salient areas in an image [7], [8], [9] or video [10], [11] by using intelligent algorithms that mimic human visual characteristics. It has been used in many image comprehension and video processing tasks, such as photo cropping [12], image editing [13], 4D saliency detection [14], photo composition [15], and target tracking [16].

With the development of deep neural networks, various network structures and novel convolution modules have been designed to improve the segmentation effect. The majority of salient object detection networks are based on the U-shaped network [17] to integrate the features of different depths and scales. The network structure represented by U-net [18] and the feature pyramid network (FPN) [19] has an obvious problem of semantic information dilution. Therefore, transferring rich semantic information to the shallow layers without losing the location information and destroying the details is the focus of current algorithms [20], [21], [22], [23], [24], [25]. Among them, the global guidance structure (GGS) represented by [24], [25], [26] is widely used. The existing methods make significant contributions to network structures and module optimizations. These designs are based on notable experimental attempts and the subjective experience of scholars. Although great progress has been made at present, two key issues still are worthy of further study, finding ways to regulate different types of features to complete better segmentation from the perspective of the overall need for weight regulation and discovering how to better balance the ability to organize a broad spatial scene with the ability to scrutinize highly detailed objects.

Most of the existing methods fuse multilevel features directly without considering the contribution ratio of the fused features to the final output. Therefore, for the first issue, we propose a perception-and-regulation (PR) block to optimize the FPN structure from the perspective of global perception and local feature fusion fine-tuning. Global perception helps to provide accurate semantic information and create better weight regulation, and local feature fusion fine-tuning helps to enhance the useful information and suppress the invalid information. The perception part of the PR block implements global perception, which adopts the deepest feature with the largest receptive field as the input. The regulation part of the PR block implements local feature fusion fine-tuning, which adopts a weighted method to optimize the feature fusion process.

For the second issue, [23], [27], [28] use atrous convolution to expand the perception range. An excessively large atrous ratio will make the information at independent sampling points discontinuous, which is not conducive to the continuity of the spatial information in detail. Therefore, the convolution combination of multiple atrous rates is widely adopted [23], [28], [29] so that the network has the ability to organize a wide spatial scene and scrutinize highly detailed objects. To solve this
We propose a perception-and-regulation (PR) block to help the network understand the global information and assign the feature weights uniformly to implement spontaneous and adaptive global feature regulation.

An imitating human eye observation module (IEO) is proposed to help the network have the ability to organize a wide spatial scene and scrutinize highly detailed objects. The PR block balances and optimizes these two abilities.

Sufficient experiments conducted on 5 SOD datasets demonstrate that the proposed method outperforms 29 state-of-the-art SOD methods in terms of eight metrics. In addition, ablation experiments on several modules and networks prove the universality of our PR block.

II. RELATED WORK

A. Salient Object Detection

Early salient object detection methods are based on handcrafted features [33], [34], [35], [36], [37], [38], [39], [40] and intrinsic cues without a deep hierarchical structure. With the development of deep learning, the deep features with rich contextual information have made great breakthroughs in the field of salient object detection. As a landmark algorithm, fully convolutional networks (FCNs) [17] creatively remove the fully connected layer to predict the semantic label for each pixel. Then, the U-shape-based structures represented by U-net [18] and the feature pyramid network (FPN) [19] have gradually become the mainstream structure by integrating all the levels of the features, layer by layer. Based on this structure, scholars have explored more abundant multilayer feature blending methods to effectively help feature expression. Among them, the global guidance structure (GGS) represented by [20], [24], [25] has gradually become a common method to strengthen the semantic information.

Our perception-and-regulation (PR) network is based on GSS to regulate the network spontaneously and adaptively. It is worth noting that, in contrast to FCN abandoning the fully connected layer (FC), our network mainly relies on the FC layer in the classification network to perceive the size and shape of objects. Then, the weights of the features to be fused are evaluated according to the semantic information obtained from the FC layer, as shown in Fig. 1.

B. Semantic Information Reinforcement

Scholars have proposed various methods to transmit the global information of high-level features to the shallower layers to help the network obtain detailed information and accurately locate the objects. For example, Hou et al. [20] proposed short connections to help shallower side-output layers obtain semantic information more directly. Zhao et al. [22] used the highest-level features to help the edge features (shallow layer) of explicit modelling filter out useless edge details. Zhao et al. [23] designed a spatial attention module where context-aware high-level features are added to help the location information transfer to the shallower layer. Liu et al. [24] introduced a global guidance module (GGM) to explicitly make shallower layers aware of the locations of the objects. The global context flow module in [25] solved the issue of dilution in the process of high-level feature transmission, which is similar to GGM.

The global guidance idea of [25] [24] can be simplified to the GGS structure in Fig. 2(a), and the structure can effectively solve the problem that the semantic information is diluted in the process of feature transfer of the FPN structure or U-net structure [18]. Even if these structural designs are based on the subjective experience and repeated attempts [20] of excellent scholars, the rigid feature fusion process does not consider the relationship between the different features to be fused and the highest level features. The GGS-PR structure of Fig. 3 uses the PR block to further slightly regulate and optimize the network. Wei et al. [21] designed a cross feature module (CFM) to select features with rich semantic information to transfer to the shallower layer and to let the features with details enter the next cycle, as shown in Fig. 2(b). CFM can be understood as strengthening the transmission of semantic information in the FPN structure. Our PR block does the same work but the method of strengthening semantic features is more flexible and adaptive, as shown in the FPN-PR of Fig. 3.

C. Attention Mechanisms

In the classification task, Hu et al. [32] greatly improves the quality of feature representation by establishing the interdependencies between the channels of its convolutional features. Correspondingly, Woo et al. [31], [41] use the spatial attention map
generated by utilizing the interspatial relationship of features with the channel attention map to help the network learn ‘where’ there is an informative part and ‘what’ is meaningful on the spatial and channel axis, respectively. Some SOD methods adopt attention mechanisms [23], [42], [43], [44]. Wang et al. [45] designed the pyramid attention module to make the network pay more attention to important regions and multiscale information. As a special attention mechanism, the gated mechanism is widely used by long short-term memory (LSTM) and gated recurrent units (GRUs), which play an important role in SOD algorithms [46], [47], [48]. Some segmentation algorithms [49], [50], [51] use the gated mechanism to adjust the network. The CGM of Unet3+ [30] can be regarded as the extreme case of the gated mechanism because the weights of the controlled features are set to 0 or 1 via argmax (Fig. 2(c)), which helps to determine whether the target is an organ or noise.

Inspired by SENet [32] and the classification network, we design a PR block for global regulation, which can be regarded as a macro global attention mechanism. [32] adaptively recalibrates channelwise features at the micro level, while the PR block recalibrates the different types of features of the whole network at the macro level. All the features to be fused are regulated in our network and the perception part of the PR block is located in the position with rich semantic information, which helps to analyse the size and shape of the object accurately and uniformly. In addition, inspired by the SAC of [29] (Fig. 2(d)), we use softmax, which is commonly used in classification networks, to add constraints to the regulation part of the PR block.

Some algorithms [23], [27], [28], [29], [51] use the atrous convolution to expand the receptive range to better observe the object. The disadvantage of an atrous convolution with a large atrous ratio is that the information given by the spatial continuity may be lost (such as edges), and it is not conducive to the segmentation of small objects. We designed a special spatial attention mechanism to compensate for this shortcoming by imitating foveal vision and peripheral vision.

III. PERCEPTION-AND-REGULATION BLOCK

The perception-and-regulation (PR) block is a computational unit with semantic perception and global regulation capability. The PR block serves the feature level weight regulation of our final network PRNet, as shown in the network structure on the left side of Fig. 3. In addition, it can also be used in many network structures or modules. The PR block adaptively adjusts the fusion process of different types of features according to the overall need for weight regulation. The SE block focuses on the features and adaptively recalibrates the features at the channel level, while our PR block focuses on the entire network and recalibrates the entire network at the feature level.
In addition, FCN [17] adapts classification networks (AlexNet [52], VGG net [53] and GoogLeNet [54]) into fully convolutional networks by replacing the fully connected (FC) layer with convolutional layers to achieve semantic segmentation. In contrast, we make full use of the perception and understanding capabilities of the FC layer to make an adaptive regulation for the entire network. Firstly, we discuss three perception strategies in detail for the perception part of the PR block. The PR block focuses on regulating different types of features to complete better segmentations from the perspective of the overall need for weight regulation. The PR block has a better regulation effect on the fusion process of features with large differences. The differences of features here refer to the different levels of features (FPN GGS), the features produced by atrous convolution or normal convolution (CFE), and the original features or attention features in the residual structure of the attention module (CBAM).

A. Perception: Semantic Information Embedding

In the perception part of the PR block, three semantic information embedding methods are designed (Fig. 4) according to the SENet idea for global information embedding. To introduce our idea clearly, we annotate the features in FPN-PR in the upper right corner of Fig. 3.

Strategy 1 in Fig. 4 uses the FC layer to perceive the size and shape of the objects. The output of the FC layer of the classification network is the category of the predicted object, while our output results are the weights of the features to be fused. The weights are adjusted adaptively according to the characteristics of the objects. Features regulated by weights are represented by the colour of the corresponding weight. Unlike SE block (upper right corner of Fig. 4) that weights multiple channels of a feature, PR block regulates multiple features in the network. The perception part of Strategy 1 is set at the location of high-level features (d5) with rich semantic information. Max pooling is used to reduce the size of the feature \( F_{m} \in \mathbb{R}^{H' \times W' \times C} \) after pooling is transformed into a one-dimensional vector \( F_{F} \in \mathbb{R}^{1 \times (H' \times W' \times C)} \) by a flatten operation. We use a multilayer perceptron (MLP) [41] with one hidden layer to enhance the perception capabilities of the network. To save parameter overhead, the hidden activation size is set to \( \mathbb{R}^{1 \times (C' \times C')}, C' \) is the number of features that need to be regulated, and it is eight for the FPN structure in Fig. 4. The output layer size is \( \mathbb{R}^{1 \times (C')}. \)

\[
P_{FC} = 2 \times (MLP_{S}(\text{Flatten}(\text{MaxPool}(F)))).
\]

The activation function of the MLPs output layer is an elementwise sigmoid.

Due to the dense connection of the FC layers, the final weights of Strategy 1 have a strong correlation. To explain the perception part of the PR block more clearly, we provide two other spatial dimension (S) and channel dimension (C) perception designs. Different from Strategy 1, we design multiple independent memory units (MUs) for Strategy 2 and 3 (Fig. 4) to better illustrate the mapping relationship. The function of MU is to establish the mapping relationship between the shape and size of the input features \( F \) and the weight of the regulated features. The number of MUs is determined by the number of features that need to be regulated. In the MU of Strategy 2, we use two convolution operations to gradually reduce the channel dimension of the input features to 1. Here, we refer to the design of the spatial attention map [31], [41]. \( r \) is the reduction ratio. We adopt the sigmoid activation function and multiply the final result by 2. In short, the perception part \( (P_{S}) \) is computed as:

\[
P_{S} = G(2 \times C_{S}(C_{S}(F))),
\]

where \( C_{S}(\cdot) \) is the convolution operation using the elementwise sigmoid function and \( G(\cdot) \) is the global average pooling.
For the MU of Strategy 3, we use global average pooling to reduce the input feature to one dimension. Then we use the fully connected layer to reduce the channel to \(C/r\). After the sigmoid and average operations, the final weight is obtained. The perception part \((P_C)\) is computed as:

\[
P_C = \text{Avg}(2 \ast FC_S(G(F))),
\]

where \(FC_S(\cdot)\) is the fully connected layer and the activation function of the output layer is sigmoid. \(\text{Avg}(\cdot)\) is the average operation of the one-dimensional vector.

### B. Regulation: Adaptive Recalibration

The accurate location information of high-level features in the FPN structure is diluted in the process of multiple fusion [25]. This is because elementwise addition or concatenation operations do not treat the weights of the features to be fused differently. Most of the current algorithms focus on changing the network structure [20], [22], [24], [25] or the modules [21], [45] to enhance the semantic information, while the PR network focuses on the regulation of the network and can greatly improve the performance of simple network structures. The PR network uses a PR block for global regulation. It builds a bridge between the features to be fused and the semantic information made by the fully connected layer. The semantic information is expressed in the form of weight.

#### 1) Basic Structural Analysis (FPN-PR and GGS-PR):

Both [24] and [25] use the global guidance method to enhance the semantic information of the shallow features of the network. We add a global guidance structure to the FPN to imitate this process, as shown in Fig. 2(a). The global guidance structure can be considered a simplified version of [20], [24], [25]. The location features are directly added to the shallow features to enhance the salient object location. There is still room for improvement. We added a PR block to both the FPN and GGS structures (Fig. 3) for further slight regulation. [21] proposes a CFM module to help the output features with high-level features (Fig. 2(b) blue arrow) as the main components to transfer to the shallow layer. This scheme is not flexible enough because the weights of features in the structure of CFM are not regulated. The PR block helps us to adaptively recalibrate the weight of each feature in the entire network according to the characteristics of the object to achieve better segmentation.

Fig. 3 shows the implementation of perception and regulation of the PR block. To simplify the network and reduce the amount of computation, we use convolution (convolution, batch normalization, ReLU) to unify the output features of the encoder to 64 channels. FPN-PR and GGS-PR on the right side of Fig. 3 do not show this detail. Taking FPN-PR with perception Strategy 2 as an example, eight memory units are used to perceive the input feature and evaluate the weights of interlayer features (i1, i2, i3, i4, i5) and decoder features (d1, d2, d3, d4, d5). The grey dotted arrow represents the input feature \(F\) of the perception part. The grey solid arrows represent the regulation of the eight output weights of the PR block on each feature to be fused. The only difference from the traditional FPN structure is that the PR block weights these features. The features \((g_1, g_2, g_3)\) in the purple region \(G\) are the global guidance features. It is worth noting that we only weighted the three feature fusion processes of the GGS to explore the influence of the PR block on the global guidance. In short, the FPN-PR and GGS-PR are computed as:

\[
d_{j-1}^{\text{FPN}} = C(W_{i_{j-1}} \ast i_{j-1} + U(W_{d_{j}} \ast d_{j})),
\]

\[
d_{j-1}^{\text{GGS}} = C(W_{i_{j-1}} \ast i_{j-1} + U(W_{d_{j}} \ast d_{j}) + U(W_{g_{j-1}} \ast g_{j-1})),
\]

where \(C(\cdot)\) refers to the convolution operation. \(W\) is the weight. \(U(\cdot)\) is an upsampling.

To make the perception part of the PR block have a better perception effect on different scale objects, we adopt the partial encoder design of the SSD algorithm [55], as shown in the left part of Fig. 3. The advantage of SSD is that it can detect objects on multiple scales. We use its 6-, 7-, and 8-layer structure and combine 4-, 7-, and 8-layer features with concatenation to realize multilayer perception.

#### 2) Exemplars (CFE-PR, CBAM-PR and EGNet-PR):

To verify the universality of the PR block, we apply it to the context-aware feature extraction module (CFE) in Fig. 2(e) [23] and the CBAM module in Fig. 2(f) [31]. The features to be fused in the FPN and GGS structures are features of different scales and depths, while the features to be fused in the CFE-PR modules have different receptive fields.

Different from [23], there are all \(3 \times 3\) convolutions in the CFE of this paper, and their dilation rates are 1, 3, 5, and 7. CFE is used in the feature i3, i4, and i5 positions (following [23]) of the basic FPN structure to enhance the output feature, and the PR block is used for internal regulation, as shown in Fig. 2(e).

As complementary attention modules, channelwise attention and spatial attention are used to calibrate features at the spatial and channel levels, and they can learn ‘what’ and ‘where’ to attend in the channel and spatial axes, respectively [31], [32], [41]. The PR block can be considered a third type of attention, which focuses on the influence of the entire feature on the network (FPN, GGS) or module (CFE, CBAM). The perception part of the PR block analyses the global context information of high-level features and strengthens or weakens the features to be fused from the perspective of the global needs of the network. Therefore, we call it global attention in Fig. 2(f). We use a PR block with channel attention and spatial attention to further optimize the attention mechanism. The regulation part of the PR block is added to the attention branch of CBAM in Fig. 2(f), and five CBAM-PR modules are added to the i1, i2, i3, i4, and i5 positions of the basic FPN structure.

We also added a PR block to the final output position of EGNet [22] for perception and regulation, the final result was further improved. Fig. 2(g) is a simplified structure diagram of EGNet, and we added a PR block in its final output location (FPN structure). Lines 11-16 of Table I prove the effect of the PR block in the above modules and networks.

### IV. Imitating Eye Observation Module

The purpose of the imitating human eye observation module (IEO) is to quickly and accurately find and locate salient objects.
The problem can be solved by partitioning and stacking PVMs in the spatial dimension \((\text{Par}_\text{sp})\) to achieve an accurate understanding of the salient objects in Step3. The peripheral vision can be considered a special attention mechanism. It expands the receptive field by comparing the features of the corresponding positions in other partitions and then corrects the original feature. In Step2, the integration operation (Fig. 6) is 3 convolution 1 in PVM and activated with ReLU.

\[
F_A, F_B, F_C, F_D = \text{Par}_\text{sp}(F_1),
\]

\[
F_A, F_B, F_C, F_D = \text{Par}_\text{sp}(F_A),
\]

\[
F^{\text{PVM}}_A = C^2_S(\text{Cat}_\text{sp}(\text{Par}_\text{ch}(C^1 \times (\text{Cat}_\text{ch}(F_A, F_B, F_C, F_D)))),
\]

\[
F'_A = F_A + F_A * F^{\text{PVM}}_A,
\]

where \(\text{Par}_\text{sp}(\cdot), \text{Par}_\text{ch}(\cdot), \text{Cat}_\text{sp}(\cdot), \text{Cat}_\text{ch}(\cdot)\) refer to spatial dimension partition, channel dimension partition, spatial dimension concatenation, channel dimension concatenation. \(C^1(\cdot) \in \mathbb{R}^{4C}\) is \(3 \times 3\) convolution 1 in PVM and activated with ReLU. \(C^2_S(\cdot) \in \mathbb{R}^1\) is \(3 \times 3\) convolution 2 in PVM and activated with sigmoid.

In Step 2, we merge the results of the partition search in the spatial dimension to obtain \(F_2\) and then use a concatenation operation to merge \(F_1\) and \(F_2\) in the channel dimension. Step 2 strengthens the association between the region search feature and the original feature.

\[
F_2 = \text{Cat}_s(F'_A, F'_B, F'_C, F'_D),
\]

\[
F_3 = C(\text{Cat}_c(F_1, F_2)),
\]

where \(C(\cdot) \in \mathbb{R}^C\).

In Step 3, we perform the peripheral visual perception of \(F_3\) again. This process is the same as the previous Eq.7-9. PVM can be considered a special attention mechanism. It expands the receptive field by comparing the features of the corresponding positions in other partitions and then corrects the original features in the spatial dimension \((\text{F}_1 \rightarrow \text{F}_2\) in Fig. 6). Eq.9 shows the combination of the attention branch and primitive branch [31], [41]. PVM can also be considered a special atrous convolution and has only four sampling points and a very large atrous rate (Fig. 6 \(F_3 \rightarrow F_3\)). The PVM here needs to be used for the feature of a large receptive field. Otherwise, the spatial information discontinuity caused by a large atrous rate will appear. The problem can be solved by partitioning and stacking PVMs \((\text{F}_1 \rightarrow \text{F}_2 \rightarrow \text{F}_3)\). The partition search strategy allows our IEO module to be used for features of the smaller receptive field,
where PVM(·) is to repeat 7-9 for F3. W1, W2, W3 are the weights produced by the memory unit. Inspired by the classification network, 13 uses a softmax layer to associate weights. SAC uses a similar approach (1-S) to associate partial weights [29]. The three features that are regulated are the peripheral vision feature, foveal vision feature (an original feature after atrous convolution operation), and original feature. The atrous ratio of the convolution of the foveal vision feature is \( r = 5 \), which is designed according to the experimental effect.

V. SUPERVISION

We use the widely used binary cross entropy (BCE) loss and a consistency-enhanced (CEL) loss [60] to supervise the prediction map, as shown in (15).

\[
L = L_{\text{bce}} + L_{\text{cel}}. \tag{15}
\]

BCE loss is defined as:

\[
L_{\text{bce}} = -\sum_{(x,y)} [g(x,y) \log(p(x,y)) + (1 - g(x,y)) \log(1 - p(x,y))], \tag{16}
\]

where \( p(x,y) \in [0, 1] \) is the prediction result of saliency map at \((x, y)\). \( g(x,y) \in [0, 1] \) is the ground truth label of the pixel \((x, y)\).

CEL loss is a variant of IoU loss, which can measure the similarity of two images from an overall perspective. It is defined as:

\[
L_{\text{cel}} = \frac{\sum_{(x,y)} [p(x,y) + g(x,y) - 2 * g(x,y) * p(x,y)]}{\sum_{(x,y)} [g(x,y) + p(x,y)]}. \tag{17}
\]

VI. EXPERIMENT

A. Datasets and Evaluation Metrics

We evaluate the proposed architecture on 5 SOD datasets: DUTS [61] with 10,553 training and 5,019 test images, DUT-OMRON [62] with 5,168 images, ECSSD [38] with 1,000 images, PASCAL-S [62] with 850 images and HKU-IS [63] with 4,447 images. We follow the data partition of [20], [60] to use 1,447 images of HKU-IS for testing.

In addition, we define a large salient object dataset (L) and a small salient object dataset (S). They are helpful for further analysing the dynamic regulation of the PR blocks when dealing with objects of different scales. We select large object images (1270) and small object images (1576) based on the ratio of the white pixels in the GT, as shown in (18) and Fig. 7. The pictures and ground truth labels are selected from five common test datasets (ECSSD, PASCAL-S, DUT-OMRON, DUTS and HKU-IS). \( P_w \) is the number of white pixels in the GT, and \( P_b \) is the number of black pixels. \( t_1, t_2 \) is the threshold. We set \( t_1 \) and \( t_2 \) to 0.38 and 0.03, respectively, to obtain datasets L and S.

\[
\text{img} \in \text{L}, \quad \text{if} \quad \frac{P_w}{P_w + P_b} > t_1
\]

\[
\text{img} \in \text{S}, \quad \text{if} \quad \frac{P_w}{P_w + P_b} < t_2. \tag{18}
\]

We use six metrics to evaluate the performance of PR-Net and the other state-of-the-art models. Mean absolute error
(MAE) [64] measures the average pixel-level relative error between the prediction and the GT by calculating the mean of the absolute value of the difference. The F-measure \(F_\beta\) [33] has also been widely adopted in previous models [20], [22]. \(F_\beta\) is the weighted harmonic mean of Precision and Recall, and \(\beta^2\) is usually set to 0.3. The maximal \(F_\beta\) values are calculated from the PR curves, represented as \(F_{\text{max}}\). An adaptive threshold (twice the mean value of the prediction) is adapted to calculate \(F_{\text{avg}}\). The weighted F-measure \(\left(F_{\text{w}}^\alpha\right)\) is a measure of completeness for improving the F-measure [65]. The structural similarity measure \(\left(S_m, \alpha = 0.5\right) [66]\) and E-measure \(\left(E_m\right) [67]\) are also useful for the quantitative evaluation of the saliency maps. In addition, precision-recall (PR) curves are drawn.

B. Implementation

We follow [21], [60], [68] to use DUTS-TR [61] as the training dataset, and the other abovementioned datasets are used as testing datasets. In the training phase, we follow [60] to use random horizontal flipping, random colour jittering, and random rotating as data augmentation techniques to prevent the overfitting problem. PRNet is trained for 40 epochs on an NVIDIA RTX 2080Ti GPU. The batch size is set to 4. VGG-16, pretrained on the ImageNet dataset, is used as the backbone network. The parameters for the rest of PRNet are initialized by the default setting of PyTorch. Our model adopts the stochastic gradient descent (SGD) optimizer with a momentum of 0.9, a weight decay of 0.0005 and an initial learning rate of 0.001. The “Poly” strategy [69] (factor is 0.9) is applied. During testing, the input size is set to 320×320.

C. Ablation Studies

**Ablation analysis of PR blocks in different structures and modules**: Lines 1–4 of Table I analyse the effect of the PR block with different perception strategies in the FPN structure. \(\text{PR}_{\text{FC}}, \text{PR}_{\text{C}}, \text{PR}_{\text{S}}\) are Strategy 1, 2, and 3 in Fig. 4. \(\text{PR}_{\text{FC}}\) has the best regulation effect in the FPN structure because of its rich parameters and intensive interaction analysis in the fully connected layer. However, in the final network structure (PRNet), \(\text{PR}_{\text{S}}\) is the best (Table IV), and we will explain this phenomenon in a later analysis. To simplify the experiment, we uniformly use the best \(\text{PR}_{\text{S}}\) in the final network to carry out the comparison experiments and the ablation experiment in Tables I, II and III. Line 6 of Table I is the gate strategy provided by GateNet [51], and its effect is not as good as that of the PR block of the global perception and regulation. Line 7 of Table I is the AIM module of MNet [60]. AIM has a more complex interaction structure than the PR block, but the regulation effect of the PR block is better. Line 8 verifies the effect of the multilevel perception strategy in the FPN structure. \(\text{FPN}_{\text{ssd}}(\text{PR}_{\text{S}})\) means that FPN-PR uses the encoder structure shown on the left side of Fig. 3. Lines 9–18 verify the improvement of the PR block to the GGS structure (the lower right corner of Fig. 3), CFE module (Fig. 2(e)), CBAM module (Fig. 2(f)), EGNet network (Fig. 2(g)) and CFD decoder (Fig. 2(b)).

**Ablation analysis of the PRNet**: Table II shows the ablation experiment of PRNet (Fig. 3). The baseline is \(\text{GGS}_{\text{ssd}}(\text{PR}_{\text{S}})\). IEO improves the network performance greatly in Line 2, but the allocation of the foveal vision features and peripheral vision features is not balanced. The performance of IEO is further improved after being regulated by the PR block (Line 3). The right side of Fig. 5 shows how the weights obtained by the perception part regulate the features of IEO. Lines 1 and 2 show the effect of multilevel perception. The experiment in the 5th line replaces the IEO-PR module in the 3 rd line with the CFE-PR module (Fig. 2(e)), which proves the effectiveness of the IEO-PR module. The model of the experiment in the 6th line is the same as that in the 3 rd line, but the experiment in the 6th line is only supervised by BCE loss (CEL loss is removed). The 3 rd and 6 th experiments show the effect of the CEL loss.

It is worth noting that, as shown in Fig. 3, the IEO and CFE modules in Table II are placed in the i3, i4 and i5 positions,
which follows the setting of [23]. The peripheral vision module in the IEO module has a large void rate, so it is better to use it for high-level features with a larger receptive field. Experiments in Table III also show that this scheme is the best for the IEO-PR module. In addition, reducing the number of IEO modules is conducive to simplifying the model and improving the speed.

Perception strategy analysis: Lines 1, 2, 3 and 4 of Table I show that Strategy 1 (PR_\text{FC}) is the best in simple structures (FPN). While Strategy 2 (PR_\text{S}) performs best in complex structures (PRNet), as shown in Table IV.

The fully connected layers in PR_\text{FC} that are inspired by the classification network make the weights coupled and correlation strong, as shown in Fig. 4. In addition, there are many parameters in the FC layer, which is also helpful to recalibrate the weights of features with obvious differences in the simple structure (FPN). However, for the complex network PRNet, the difference between the features to be fused decreases. Taking the feature fusion process F_\text{G} on the far right of PRNet (Fig. 3) as an example, because d4 is the fusion feature of g2 (d5) and i4, the difference between g2 and d4 is reduced. The fully connected layers in Strategy 1 overinterpret the weights, which worsens the effect of the PR_\text{FC} block. The fully connected layers are also used in Strategy 3 (PR_\text{C}), so it has the same problem. Strategy 1 performs better on simple networks (FPN, CFD), and Strategy 2 performs better on complex networks (PRNet, EGNNet). In addition, we analyse the weights of Strategies 1 and 2 in the FPN structure, as shown in Fig. 8. We can find that Strategy 1 is radical and sensitive, while Strategy 2 is conservative and restricted.

These characteristics result in the performance differences between the two strategies in the simple network structures and complex network structures.

PR_\text{S} directly uses global average pooling to reduce the spatial dimension (H, W) to (1, 1), which is beneficial for complex networks (PRNet) with very little difference in the features to be fused. Because there is no fully connected layer, the final weights are more directly and closely related to the spatial features of salient objects [70]. According to the above analysis, we can use Strategy 1 to regulate the feature fusion process with large feature differences, while Strategy 2 can be used for the feature fusion process with small feature differences.

Feature weight analysis: To further analyse how the PR block works, we show the average weight of each fusion position in the pie chart and line chart in Fig. 9. F1-4 and F_\text{G}1-3 represent multiple regulated points of FPN-PR and GGS-PR (Fig. 3). The pie chart shows the results in the training dataset and 5 test datasets. The line chart shows the average weight of multiple fusion position weights of 5 test datasets and the training dataset. In the line chart, we use 1270 images of large objects and 1576 images of small objects to show how the PR block regulates the network in the segmentation task of objects of different scales.

We use the dotted circle to indicate the position where the weight of the PR block obviously changes. This change is related to the size of the salient object. It is worth noting that the highest-level features (d5) of the small objects are specifically enhanced (dotted circle in the small object) to prevent dilution. The lowest-level (i1) features in the large objects are sufficiently suppressed to prevent interference. The lower part of Fig. 9 shows the effect of the PR block. To verify that the dynamic regulation of weights...

Fig. 8. The effect analysis of Strategy 1 (PR_{FC}) and Strategy 2 (PR_{S}) in the FPN architecture. The pie chart and line chart show the statistical results of the feature weights on five test datasets. The pie chart is presented in the proportion of the feature weight. The line chart is presented in the form of a concrete numerical value. F1-F5 represent the five feature fusion processes of FPN. F1 represents the feature fusion process of i1 and d2. The blue regions represent the weight ratio of the decoder features (d2, d3, d4, d5). The pink regions represent the weight ratio of the interlayer features (i1, i2, i3, i4). For the feature definitions of i1 and d2, please refer to FPN-PR in the upper right corner of Fig. 3.

Fig. 9. Weight analysis. For FPN-PR_\text{S} and GGS-PR_\text{S}, we use pie charts and line charts to show the average values of multiple fusion position weights of 5 test datasets and the training dataset. In the line chart, we use 1270 images of large objects and 1576 images of small objects to show how the PR block regulates the network in the segmentation task of objects of different scales.
is meaningful, we lock the weights in Table I Line 5. The weights of FPN(PR\textsubscript{S}-fixed) are obtained from the average weight of the training dataset. Lines 1, 4 and 5 show that it is effective to suppress low-level features with fixed weights, but it is better to regulate the weights according to the analysis results of the PR block.

D. Comparison With State-of-The-Art

PRNet is mainly designed for VGG-16. To verify the effectiveness of the method, we also design a model based on the ResNet-50 backbone. For the convenience of distinguishing, we use PRNet\textsubscript{V} and PRNet\textsubscript{R} to represent the VGG-16-based PRNet and the ResNet-50-based PRNet. Considering the different depths of the two types of backbones and the particularity of the GGS structure, we slightly adjust the structure of PRNet\textsuperscript{R}. Please refer to the supplementary materials for specific details.

We compare PRNet against 29 SOD state-of-the-art algorithms. Algorithms based on the VGG-16 backbone network include DCL [71], NLDF [72], MSRNet [73], DSS [20], BMPM [50], RAS [42], PAGRN [43], C2S [74], PAGE [45], BANet [75], AFNet [57], GateNet [51], ITSD [76], CAGNet [77], etc. Algorithms based on lightweight backbone networks include FCNet [78], HVPNet [79], SAMNet [80], etc. Algorithms based on the ResNet-50 backbone include PiCANet [81], R3Net [82], CapSal [48], CPD [44], BASNet [83], etc. The comparison algorithms also include U\textsuperscript{2}-Net [84] without a backbone.

For fair comparisons, we use all the saliency maps provided by the authors or generated by their codes. PoolNet [24] (with StdEdge) adds another dataset (BSDS500) for joint training, which makes the comparison results unfair. Therefore, we compare PRNet with PoolNet (with SalEdge, only the DUTS-TR training dataset) in Table V. The experimental result shows that our algorithm is better. Experiments show that the performance of U\textsuperscript{2}-Net is close to that of PRNet\textsuperscript{R}. Comprehensively comparing speed, model size and performance, PRNet\textsubscript{V} and PRNet\textsubscript{R} are better.

Quantitative evaluation: Table V shows the scores of the proposed model and 29 state-of-the-art saliency detection methods on five widely used datasets and also demonstrates that the perception-and-regulation strategy is successful in making simple networks perform favourably against other algorithms. Moreover, the PR curves by our approach outperform the other methods, as shown in Fig. 10.

PRNet also has obvious advantages in model size, speed and computational complexity. The proposed PRNet\textsubscript{V} and PRNet\textsubscript{R} are simple networks, so there is no comparison with large parameter models. The model sizes of PRNet\textsubscript{V} and PRNet\textsubscript{R} are 27.57 M and 35.85 M. They are smaller than most previous algorithms (U\textsuperscript{2}-Net (176.3 M), PoolNet (66.66 M), AFNet (143 M), MLMSNet (263 M), BANet\textsubscript{V} (55.9 M), HRS (129.6 M), GateNet\textsubscript{V} (125.62 M), etc.). The computational complexity of PRNet\textsubscript{V} and PRNet\textsubscript{R} is 48.76 GMac and 21.0 GMac, which are better than the comparison algorithms (NLDF (263.9 GMac), UCF (61.4 GMac), DSS (114.6 GMac), PoolNet (194.27 GMac), ITSD (181.65 GMac), GateNet\textsubscript{V} (228.07 GMac), etc.).

The speeds of PRNet\textsubscript{V} and PRNet\textsubscript{R} are 74 FPS and 53 FPS.

Qualitative evaluation: Fig. 11 shows the visual examples produced by our model and other models. From the 1st row to the 10th row, the size of the salient object gradually changes from the largest to the smallest. Our algorithm is effective in dealing with multiscale object detection. The proposed method is competitive with state-of-the-art detection algorithms.
Fig. 10. Precision-Recall curves (1st row) and F-measure curves (2nd row) on five common saliency datasets. PRNet\(^V\) was used for comparison.

Fig. 11. Qualitative comparisons with the state-of-the-art algorithms. From top to bottom, the size of salient objects gradually decreases.

Fig. 12. Visual analysis of the PR block regulation process. We show two examples (more examples can be found in the supporting document). For the feature analysis of the bird, the 1st and 2nd rows are the decoder features d1-d5 of FPN and the decoder features d1-d5 of FPN-PR\(_FC\), respectively. The 3rd row is ground truth, the input image and encoder (VGG-16) features i1-i4. The 1st column is the final output saliency maps of FPN and FPN-PR\(_FC\). In the red box is the feature fusion process of FPN-PR\(_FC\). The grey (encoder), red (interlayer features), and blue (decoder features) arrows correspond to the FPN-PR\(_FC\) in the upper right corner of Fig. 3. The white font is the result of weight regulated by PR block (PR\(_FC\)).
performs better in various challenging scenarios, including the small, medium-sized and large objects. Fig. 12 shows the visualization results of the whole process. The difference between the FPN network with a PR block and the FPN network without a PR block is clearly shown. In addition, we provide some failure cases of our algorithm in the supporting document to help future researchers conduct further analysis.

VII. CONCLUSION

In this paper, we propose a novel framework, PRNet, for salient object detection. A PR block is designed to help the network understand the global information and assign the feature weights spontaneously and adaptively. To better perceive semantic information and reasonably allocate weight, we propose multiple perception strategies and carry out comparative experiments. Considering the relationship between local perception and global perception, we propose an IEO module so the network has the ability to organize a wide spatial scene and to scrutinize highly detailed objects. Sufficient experiments demonstrate that PRNet performs well. In the future, we may expand PRNet to more complex structures, such as recurrent structure networks and multimodal SOD networks (RGB-T, RGB-D).

REFERENCES

[1] W. Wang et al., “Salient object detection in the deep learning era: An in-depth survey,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 6, pp. 3299–3329, Jun. 2022.
[2] B. Jiang, Z. Zhou, X. Wang, J. Tang, and B. Luo, “CMSalGAN: RGB-D salient object detection with cross-view generative adversarial networks,” IEEE Trans. Multimedia, vol. 23, pp. 1343–1353, 2021.
[3] D.-P. Fan et al., “Re-thinking co-salient object detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 8, pp. 4339–4354, Aug. 2022.
[4] D.-P. Fan, Z. Lin, Z. Zhang, M. Zhu, and M.-M. Cheng, “Rethinking RGB-D salient object detection models, data sets, and large-scale benchmarks,” IEEE Trans. Neural Netw. Learn. Syst., vol. 32, no. 5, pp. 2075–2089, May 2021.
[5] J. Zhang et al., “Learning saliency from single noisy labelling: A robust model fitting perspective,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 43, no. 8, pp. 2866–2873, Aug. 2021.
[6] Q. Chen et al., “EFNet: A novel enhancement and fusion network for RGB-D saliency detection,” Pattern Recognit., vol. 112, 2021, Art. no. 107740.
[7] J. Zhang et al., “Weakly-supervised salient object detection via scribble annotations,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 12543–12552.
[8] Q. Chen et al., “RGB-D salient object detection via 3d convolutional neural network model,” in Proc. AAAI Conf. Artif. Intell., 2021, pp. 1063–1071.
[9] J. Zhang et al., “Uncertainty inspired RGB-D saliency detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 44, no. 9, pp. 5761–5779, Sep. 2022.
[10] W. Wang, J. Shen, and L. Shao, “Video salient object detection via fully convolutional networks,” IEEE Trans. Image Process., vol. 27, no. 1, pp. 38–49, Jan. 2018.
[11] W. Wang, J. Shen, R. Yang, and F. Porikli, “Saliency-aware video object segmentation,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 40, no. 1, pp. 20–33, Jan. 2018.
[12] W. Wang, J. Shen, and H. Ling, “A deep network solution for attention and aesthetics aware photo cropping,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 7, pp. 1531–1544, Jul. 2019.
[13] M.-M. Cheng, F.-L. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, “Repfinder: Finding approximately repeated scene elements for image editing,” ACM Trans. Graph., vol. 29, no. 4, pp. 1–8, Jul. 2010.
[14] T. Wang, Y. Piao, H. Lu, X. Li, and L. Zhang, “Deep learning for light field saliency detection,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 8837–8847.
[15] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, “Sketch2photo: Internet image montage,” ACM Trans. Graph., vol. 28, no. 5, pp. 1–10, Dec. 2009.
[16] V. Mahadevan and N. Vasconcelos, “Saliency-based discriminant tracking,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 1007–1013.
[17] J. Long, E. Shelhamer, and T. Darrell, “Fully convolutional networks for semantic segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 3431–3440.
[18] O. Ronneberger, P. Fischer, and T. Brox, “U-Net: Convolutional networks for biomedical image segmentation,” in Proc. Med. Image Comput. Assis. Interven., vol. 9351, 2015, pp. 234–241.
[19] T.-Y. Lin et al., “Feature pyramid networks for object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 2117–2125.
[20] Q. Hou et al., “Deeply supervised salient object detection with short connections,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 41, no. 4, pp. 815–828, Apr. 2019.
[21] J. Wei, S. Wang, and Q. Huang, “F 3 Net: Fusion, feedback and focus for salient object detection,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 12321–12328.
[22] J. Zhao et al., “EgNet: Edge guidance network for salient object detection,” in Proc. IEEE Int. Conf. Comput. Vis., 2019, pp. 8778–8787.
[23] T. Zhao and X. Wu, “Pyramid feature attention network for saliency detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 3080–3089.
[24] J.-J. Liu, Q. Hou, M.-M. Cheng, J. Feng, and J. Jiang, “A simple pooling-based design for real-time salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 3912–3921.
[25] Z. Chen, Q. Xu, R. Cong, and Q. Huang, “Global context-aware progressive aggregation network for salient object detection,” in Proc. AAAI Conf. Artif. Intell., 2020, pp. 10599–10606.
[26] S. Yu, B. Zhang, J. Xiao, and E. G. Lim, “Structure-consistent weakly supervised salient object detection with local saliency coherence,” in Proc. AAAI Conf. Artif. Intell., 2021, pp. 3234–3242.
[27] D.-P. Fan et al., “Camouflaged object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 2774–2784.
[28] 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 Trans. Pattern Anal. Mach. Intell., vol. 40, no. 4, pp. 834–848, Apr. 2018.
[29] S. Qin, L. C. Chen, and A. Yuille, “Detectors: Detecting objects with recursive feature pyramid and switchable atrous convolution,” in Proc. IEEE Int. Conf. Comput. Vis., 2021, pp. 10208–10219.
[30] H. Huang et al., “UNet3: A full-scale connected unet for medical image segmentation,” in Proc. IEEE Int. Conf. Acoust., Speech Signal Process., 2020, pp. 1055–1059.
[31] S. Wso, J. Park, J. Lee, and I. S. Kweon, “CBAM: Convolutional block attention module,” in Proc. Eur. Conf. Comput. Vis., 2018, no. 1121, pp. 3–19.
[32] J. Hu, L. Shen, S. Albanie, G. Sun, and E. Wu, “Squeeze-and-excitation networks,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 8, pp. 2011–2023, Aug. 2020.
[33] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, “Frequency-tuned salient region detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2009, pp. 1597–1604.
[34] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. S. Torr, and S.-M. Hu, “Global contrast based salient region detection,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 37, no. 3, pp. 569–582, Mar. 2015.
[35] Z. Jiang and L. S. Davis, “Submodular salient region detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2013, pp. 2043–2050.
[36] D. A. Klein and S. Frintrop, “Center-surround divergence of feature statistics for salient object detection,” in Proc. Int. Conf. Comput. Vis., 2011, pp. 2214–2221.
[37] H. Jiang et al., “Salient object detection: A discriminative regional feature integration approach,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2013, pp. 2083–2090.
[38] Q. Yan, L. Xu, J. Shi, and J. Jia, “Hierarchical saliency detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2013, pp. 3166–3173.
[39] W. Wang, J. Shen, L. Shao, and F. Porikli, “Correspondence driven saliency transfer,” IEEE Trans. Image Process., vol. 25, no. 11, pp. 5025–5034, Nov. 2016.
ZHU et al.: PERCEPTION-AND-REGULATION NETWORK FOR SALIENT OBJECT DETECTION

[41] J. Park, S. Woo, J. Lee, and I. S. Kweon, “BA-M: Bottleneck attention module,” in Proc. Brit. Mach. Vis. Conf., 2018, Art. no. 147.

[42] S. Chen et al., “Reverse attention based residual network for salient object detection,” IEEE Trans. Image Process., vol. 29, pp. 3763–3776, 2020.

[43] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, “Progressive attention guided recurrent network for salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 714–722.

[44] Z. Wu, L. Su, and Q. Huang, “Cascaded partial decoder for fast and accurate salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 3902–3911.

[45] W. Wang, S. Zhao, J. Shen, S. C. H. Hoi, and A. Borji, “Salient object detection with pyramid attention and salient edges,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 1448–1457.

[46] W. Wang, J. Shen, M.-M. Cheng, and L. Shao, “An iterative and cooperative top-down and bottom-up inference network for salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 5961–5970.

[47] W. Wang, J. Shen, X. Dong, A. Borji, and R. Yang, “Inferring salient objects from human fixations,” IEEE Trans. Pattern Anal. Mach. Intell., vol. 42, no. 8, pp. 1913–1927, Aug. 2020.

[48] L. Zhang, J. Zhang, Z. Lin, H. Lu, and Y. He, “Capsal: Leveraging captioning to boost semantics for salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 6017–6026.

[49] M. A. Islam, M. Rochan, N. D. B. Bruce, and Y. Wang, “Gated feedback refinement network for dense image labeling,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2017, pp. 4877–4885.

[50] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, “A Bi-directional message passing model for salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 1741–1750.

[51] X. Zhao, Y. Pang, L. Zhang, H. Lu, and L. Zhang, “Suppress and balance: A simple gated network for salient object detection,” in Proc. Eur. Conf. Comput. Vis., 2020, pp. 35–51.

[52] A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Proc. Conf. Neural Inf. Process. Syst., 2012, pp. 1106–1114.

[53] K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Proc. Int. Conf. Learn. Representations, 2015, pp. 1–14.

[54] C. Szegedy et al., “Going deeper with convolutions,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 1–9.

[55] W. Liu et al., “SSD: Single shot multibox detector,” in Proc. Eur. Conf. Comput. Vis., 2016, vol. 9905, pp. 21–37.

[56] E. M. S. M. V. A. C. Schütz, “A review of interactions between peripheral and foveal vision,” J. Vis., vol. 20, no. 12, 2020, Art. no. 2.

[57] M. Feng, H. Lu, and E. Ding, “Attentive feedback network for boundary-aware salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 1623–1632.

[58] H. Strasburger, I. Rentschler, and M. Jüttermann, “Peripheral vision and pattern recognition: A review,” J. Vis., vol. 11, no. 5, 2011, Art. no. 13.

[59] C. Eymond, T. S. Malkinson, and L. Naccache, “Learning to see the ebbing-haus illusion in the periphery reveals a top-down stabilization of size perception across the visual field,” Sci. Rep., vol. 10, no. 1, pp. 1–10, 2020.

[60] Y. Pang, X. Zhao, L. Zhang, and H. Lu, “Multi-scale interactive network for salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 9410–9419.

[61] L. Wang et al., “Learning to detect salient objects with image-level supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2017, pp. 3796–3805.

[62] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, “The secrets of salient object segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 280–287.

[63] G. Li and Y. Yu, “Visual saliency based on multiscale deep features,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2015, pp. 5455–5463.

[64] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, “Saliency filters: Contrast based filtering for salient region detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2012, pp. 733–740.

[65] R. Margolin, L. Zelnik-Manor, and A. Tal, “How to evaluate foreground maps,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2014, pp. 248–255.

[66] D.-P. Fan, M.-M. Cheng, Y. Liu, T. Li, and A. Borji, “Structure-measure: A new way to evaluate foreground maps,” in Proc. IEEE Int. Conf. Comput. Vis., 2017, pp. 4558–4567.

[67] D. Fan et al., “Enhanced-alignment measure for binary foreground map evaluation,” in Proc. Joint Conf. Artif. Intell., 2018, pp. 698–704.

[68] R. Wu et al., “A mutual learning method for salient object detection with intertwined multi-supervision,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 8142–8151.

[69] W. Liu, A. Rabinovich, and A. C. Berg, “ParseNet: Looking wider to see better,” Comput. Sci., 2015.

[70] M. Lin, Q. Chen, and S. Yan, “Network in network,” in Proc. Int. Conf. Learn. Representations, 2014.

[71] G. Li and Y. Yu, “Deep contrast learning for salient object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2016, pp. 478–487.

[72] Z. Luo et al., “Non-local deep features for salient object detection,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 6593–6601.

[73] G. Li, Y. Xie, L. Lin, and Y. Yu, “Instance-level salient object segmentation,” in Proc. IEEE Conf. Comput. Vis. Pattern Recognit., 2017, pp. 247–256.

[74] X. Li, F. Yang, H. Cheng, W. Liu, and D. Shen, “Contour knowledge transfer for salient object detection,” in Proc. Eur. Conf. Comput. Vis., 2018, pp. 370–385.

[75] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, “Selectivity or invariance: Boundary-aware salient object detection,” in Proc. Int. Conf. Comput. Vis., 2019, pp. 3798–3807.

[76] H. Zhou, X. Xie, J.-H. Lai, Z. Chen, and L. Yang, “Interactive two-stream decoder for accurate and fast saliency detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2020, pp. 9138–9147.

[77] S. Mohammadi, M. Noori, A. Bahri, S. G. Majelan, and M. Havaei, “CAGNet: Content-aware guidance for salient object detection,” Pattern Recognit., vol. 103, 2020, Art. no. 107303.

[78] D. Zhang, H. Tian, and J. Han, “Few-cost salient object detection with adversarial-paced learning,” in Proc. Conf. Neural Inf. Process. Syst., 2020, pp. 12236–12247.

[79] Y. Liu, X.-Y. Gu, X.-Y. Zhang, W. Wang, and M.-M. Cheng, “Lightweight salient object detection via hierarchical visual perception learning,” IEEE Trans. Cybern., vol. 51, no. 9, pp. 4439–4449, Sep. 2021.

[80] Y. Liu, X.-Y. Zhang, J.-W. Bian, L. Zhang, and M.-M. Cheng, “SAMNet: Stereoscopically attentive multi-scale network for lightweight salient object detection,” IEEE Trans. Image Process., vol. 30, pp. 3804–3814, 2021.

[81] N. Liu, J. Han, and M.-H. Yang, “PiCANET: Learning pixel-wise contextual attention for saliency detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2018, pp. 3089–3098.

[82] Z. Deng et al., “R 2Net: Recurrent residual network for saliency detection,” in Proc. Joint Conf. Artif. Intell., 2018, pp. 684–690.

[83] X. Qin et al., “BasNet: Boundary-aware salient object detection,” in Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit., 2019, pp. 7479–7489.

[84] X. Qin et al., “U2Net: Going deeper with nested U-structure for salient object detection,” Pattern Recognit., vol. 106, 2020, Art. no. 107404.