SBN: Scale Balance Network for Accurate Salient Object Detection

Zhenshan Tan
Department of Electronic Engineering
Fudan University
Shanghai 200433, China
zstan19@fudan.edu.cn

Xiaodong Gu
Department of Electronic Engineering
Fudan University
Shanghai 200433, China
xdgu@fudan.edu.cn

***Abstract*—Recent great progress has been made on Salient Object Detection (SOD) by deep Convolutional Neural Networks (CNNs). However, most SOD methods still suffer from scale imbalance issue, which pays more attention on large salient areas but ignores small salient areas though they belong to the same object. To address this issue, this paper proposes a Scale Balance Network (SBN) to accurately locate large salient areas and recognize small salient areas. Firstly, a backbone network specifically designed for object detection is adopted in this paper, which captures larger resolution with more spatial features in deeper layers. Secondly, to focus on the balance between the large salient areas and the small salient areas, this paper proposes a novel Connective Feature Pyramid Module (CFPM) for sufficiently leveraging the multi-scale features and the multi-level features, which includes Feature Coherence Enhancement (FCE) and Feature Progressive Extraction (FPE). FCE is designed to enhance the coherence between high-level and low-level features, and FPE is designed to extract the progressive features in different convolutional layers. Finally, an Edge Enhancement Architecture with Various Kernels (EEAVK) is proposed to refine the edge features. Experimental results on five benchmark datasets show that the proposed method outperforms or achieves consistently superior performance in comparison with other methods under different evaluation metrics.**

***Index Terms*—deep learning, salient object detection, scale balance, edge enhancement**

I. Introduction

Salient object detection aims to find the most important objects in a natural image. It has various applications on many visual tasks such as object detection, image captioning and image retrieval. Currently, most of the state-of-the-art saliency detection methods are based on deep learning models, which extract high-level context features better than traditional unsupervised stimuli-driven methods. However, there still two key problems need to be solved further. On the one hand, the interiors of the salient object may have various appearances with different scales, so that only part of the whole object can be detected. On the other hand, compared with the mainstay of the salient object, the edges fail to minimize false positives in the high texture background region.

Many researchers have made efforts to address the above two issues. Early researches are mainly based on additional hand-crafted models. For example, super-pixels methods [1] [2] are added to simplify the undesirable details. However, super-pixels focus on low-level spatial features but neglect the high-level context features, which may lose track of some important information. In addition, the method of enforcing spatial coherence with a Conditional Random Field (CRF) [3] is also added to refine the results. Although postprocessing methods can make the salient map better, the extra processing steps are time-consuming.

Besides additional hand-crafted models, attention modules methods and feature enhancement methods are introduced to the saliency detection networks. Attention modules methods generate an attention map via embedding contextual attention mechanism, which include channel-wise attention [4], pixel-wise attention [5] and pyramid feature attention [6]. Attention mechanism helps the deep networks pay more attention on the affinities of context features and refine the local region. Nevertheless, the methods lack the ability to balance different scale areas, which may excessive focus on the more salient area such as large salient areas but ignores little salient areas. Feature enhancement methods adopt a backbone feature extractor such as ResNet [7] or VGGNet [8] with pretrained weights from ImageNet, and involve extra stages to handle the objects with various scales. Typical representatives are edge enhancement models such as [9] [10] [11]. The methods attempt to leverage the edge features to locate objects in an image, especially their boundaries more accurately. However, the backbone of ResNet or VGGNet is specially designed for image classification, which adopts large down-sampling factor to recognize the category of the object instances but ignores of spatially locating the position.

To overcome the above issues, this paper proposes a novel salient object detection method named Scale Balance Network (SBN). In consideration of the gap between the image classification and object detection, a backbone network specifically designed for object detection called DetNet [12] is adopted first. Different from the original DetNet, we abandon one of the successive down-sampling but only retain the operation of maintaining spatial resolution and enlarging receptive field in deeper layers. Therefore, the backbone is able to locate the large objects more accurately and find the missing small objects. Furthermore, a novel connective feature pyramid module

This work was supported in part by National Natural Science Foundation of China under grants 61771145 and 61371148.

(CFPM) is proposed for balancing the weights between large scale salient areas and small ones. Different from existing pyramid feature attention network [6] [11] or U-net structure [13], CFPM not only focus on low-level features and high-level features, but also consider the coherence between low-level features and high-level features. In addition, in order to refine the boundaries of salient areas, an edge enhancement architecture with various kernels (EEAVK) is fused in low-level features. Different from previous researches [9] [10] [11], this paper proposed various kernels setting instead of single kernel setting for better edge detection. Benefitting from the ways of various kernels setting, EEAVK can extract more sufficient shape details and complex various edge features.

In short, the main contributions of this work can be highlighted as follows:

- This paper adopts an effective backbone network designed for object detection (DetNet) to obtain higher spatial resolutions and larger receptive fields.
- A Connective Feature Pyramid Module (CFPM) is proposed for sufficiently extracting the context information in high-level features and accurately locating the salient objects in low-level features by Feature Progressive Extraction (FPE). The coherence between high-level features and low-level features is also considered in CFPM by Feature Coherence Enhancement (FCE).
- A novel Edge Enhancement Architecture with Various Kernels (EEAVK) is proposed in this paper to learn more accurate and subtle information of boundary features.
- Without any bells and whistles, the Scale Balance Network (SBN) proposed in this paper achieves new state-of-the-art results on several benchmark datesets.

II. Related Works

Over the past two decades, hundreds of salient object detection methods have been proposed. Earlier methods estimate the salient maps using prior knowledge by hand-crafted features such as color contrast [14], boundary background [15], center prior [16] and so on. However, these methods only focus on the low-level features and local texture information, which lack the essential high-level context features.

In recent years, the methods based on Convolution Neural Networks (CNNs) have made breakthroughs in saliency detection. Early CNNs-based methods [17] [18] leverage deep learning networks to generate the feature maps to calculate saliency of image units. According to the image units, the final salient map is generated by some extra-added algorithms such as weighted sum and distance information. However, these methods are easily limited by the performance of image units. To refine the results, some pre-processing algorithms such as super-pixels methods [2] and pro-processing algorithms such as Conditional Random Field (CRF) [3] [19] are combined to the CNNs though they are time-consuming.

After then, end-to-end deep learning networks become the mainstream frameworks. Long et al. [20] firstly propose a fully convolutional networks to predict the salient labels for each pixel. Similar to their work, large numbers of pixel-wise CNNs methods such as [20] [22] [23] are proposed. In this two years, attention modules and feature enhancement have been proved useful. In paper [6], Pyramid Feature Attention Network(PFAN) is applied to extract high-level features and low-level features. Another paper [24] proposes an output-guided attention module with multiscale outputs instead of applying the widely used self-attention module. In paper [17], a progressive attention driven framework enhanced by multi-path recurrent feedback is proposed, which intrinsically refines the entire network through global semantic information from the top convolutional layer. The paper [25] proposes reverse attention to guide side-output residual learning in a top-down manner. Though the above methods improve the quality of significant results by extracting high-level and low-level features, they ignore the coherence between deeper convolutional layers and shallower convolutional layers. Moreover, these methods mainly focus on the high-level features in the deeper convolutional layers, which lack enough spatial information and scale balance information because of adopting large down-sampling factor.

Meanwhile, feature enhancement module is another effective model that attracts a lot of attention. In paper [9], the authors propose an edge-guided non-local fully convolutional neural network, which use the edge enhancement module to generate the edge guided feature for accurate salient object detection. To further integrate with FCNs and jointly optimize through end-to-end training, Wu et al. [26] present a deep guided filtering network for pixel-wise image prediction.

Compared with previous researches with attention module [5] [6] [24] [25], a novel scale balance network is proposed in this paper. SBN first maintains the spatial resolution and enlarges the receptive field by adopting a backbone that specially designed for object detection, aiming to find smaller objects. Then different from previous researches, a novel connective feature pyramid module is proposed for enhancing the coherence by feature coherence enhancement and balancing the scale by feature progressive extraction. Different from previous researches with feature enhancement module [9] [10] [26], an edge enhancement architecture with various kernels is designed to refine the edge feature, which focus on the various scale problem so that detects the boundary of salient objects better. The experimental results verify our statement.

III. Scale Balance Network

In this paper, we propose a novel scale balance network, which contains a backbone module to capture the larger spatial features to detect smaller objects, a Connective Feature Pyramid Module (CFPM) to enchance the coherence and balance the scale between high-level and low-level features by Eeature Coherence Enhancement (FCE) and feature Progressive Extraction (FPE), an Edge Enhancement Architecture with Various Kernels (EEAVK) module to refine the edges of salient objects. The overall architecture is shown in Fig. 1.

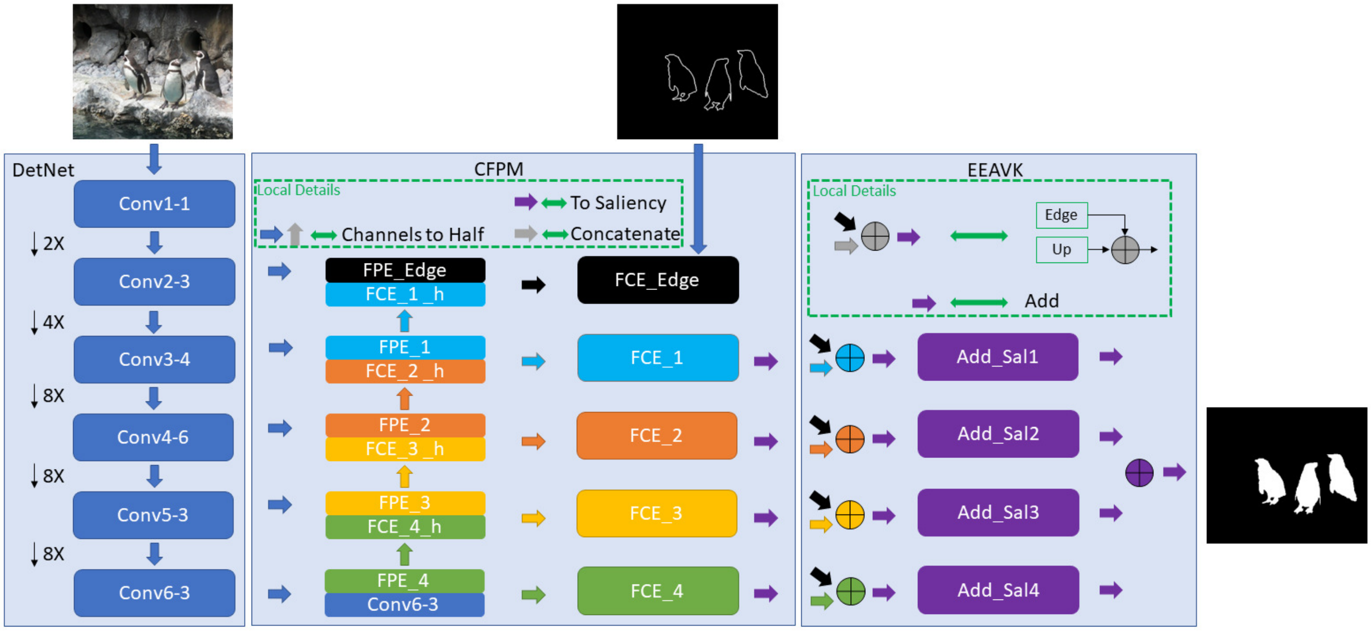

Fig. 1. Overall Architecture.CFPM: connective feature pyramid module. EEAVK: edge enhancement architecture with various kernels. Conv: Convolutional layers. More details about CFPM are shown in Fig. 2. FCE: Feature Coherence Enhancement. FPE: Feature Progressive Extraction. FCE_h: Reduce the channels to half of FCE.

A. Backbone Network

Previous researches use a backbone network such as ResNet [7] or VGGNet [8] with pre-trained weights on the ImageNet classification dataset. The aim of classification is to recognize the category of the objects while the aim of object detection is to spatially localize the bounding-boxes. The large down-sampling factors in VGGNet and ResNet can help the networks recognize the objects fast and accurately. However, the over-large down-sampling factor easily losses important spatial information especially in deeper convolutional layers, which brings the negative influence into saliency detection.

To this end, this paper adopts the backbone called DetNet [12] that specially designed for object detection. On the one hand, DetNet has exactly the same number of stages as the detector used, so additional stages can be pre-trained in the ImageNet dataset. On the other hand, DetNet maintains larger spatial resolution than other backbone networks, which is more powerful in locating the boundary of large objects and finding the missing small objects. To obtain more sufficiently boundary features, we change the beginning stride of 4 down-sampling to 2 for larger receive fields, which is shown at the DetNet part in Fig. 1.

B. Connective Feature Pyramid Module

Generally, shallow layers pay attention to the texture of objects while deep layers pay attention to the context of images, which means shallow layers have smaller receptive fields and deep layers have larger receptive fields. More importantly, different layers should have the potential connections [27]. However, most existing methods only focus on one of them but ignore the connective and distinctive between different layers, leading to the wrong prediction. Moreover, previous researches usually define the first two modules as the shallow layers and the last two or three modules as the deep layers [6] [28], which lack the interpretability. Therefore, connective feature pyramid module is designed to extract features in different convolutional layers, aiming to enhance coherence among different layers and obtain multi-scale multi-receptive-field features. In addition, edge supervision is also adopted to constrain the boundary features. As shown in Fig. 2, different from previous researches [6] [12] [29], different layers have the operations with directly concatenating and connectively passed. Different from previous researches [12] [13] [30], CFPM do not distinct the shallow layers and deep layers specifically but extract different layers features progressively by gradually rising dilation convolution [31].

a) Feature Coherence Enhancement: As shown in Fig. 2, to enhance the feature coherence between different layers, FCE adopts more connection paths that include lateral transfer path and upward transfer path. Lateral transfer path generates cross layer connection, aiming to enhance the coherence between non-nearest-neighbor layers. That is because features in deep convolutions lack low-level features and spatial information, which cannot distinct large-scale objects and precise structural edges. Upward transfer path generates adjacent layer connection, aiming to enhance the coherence between nearest-neighbor layers. That is because the features of adjacent layers are similar, and adjacent layer connection can enhance the expression of features. During the two paths, channels are divided into half to keep the sum channels unchanged.

We take Conv2-3, Conv3-4, Conv4-6, Conv5-3, Conv6-3 of DetNet to extract multi-scale features first. Conv1-1 is thrown away because the convolution is too close to the input and the receptive field is too small, and Conv2-3 is designed for edge

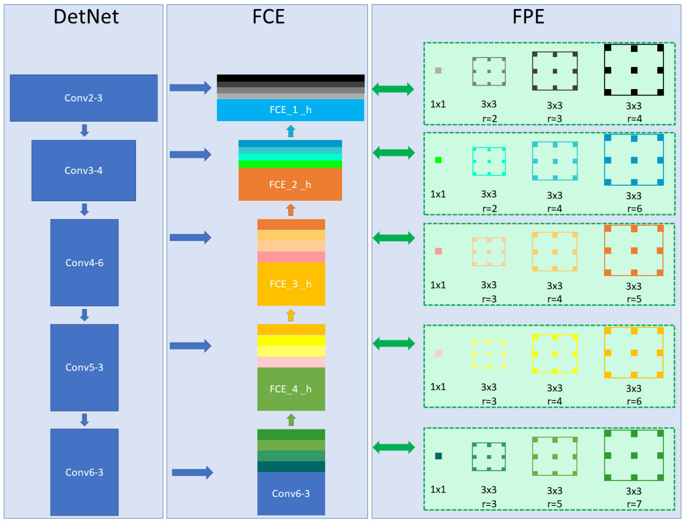

Fig. 2. Connective Feature Pyramid Module. FCE: Feature Coherence Enhancement. FPE: Feature Progressive Extraction. Conv: Convolutional layers. 1×1: the kernel size of convolution is 1×1. 3×3: the kernel size of convolution is 3×3. r: the rate of dilated convolution.

extraction [23]. The above convolutions can be represented as $\boldsymbol{f}^{D} \in \mathbb{R}^{W\times H\times C}$, where $\mathbb{R}$ denotes the convolution and W,H,C denotes width, height and channel number of each $\mathbb{R}$ respectively. Then, the five side paths in FCE can be denoted as $\boldsymbol{f}^{FCE} \in \mathbb{R}^{W\times H\times C}$. According to the lateral transfer path, the channels of $\boldsymbol{f}^{D}$ are reduced to half, which are used for feature progressive extraction and are denoted as $\boldsymbol{f}^{FPE} \in \mathbb{R}^{W\times H\times \frac{C}{2}}$. According to the upward transfer path, the channels of $\boldsymbol{f}^{FCE}$ are reduced to half, which are used for feature coherence enhancement and are denoted as $\boldsymbol{f}^{FCE_h} \in \mathbb{R}^{W\times H\times \frac{C}{2}}$. Lastly, $\boldsymbol{f}^{FPE}$ and $\boldsymbol{f}^{FCE_h}$ are concatenated to $\boldsymbol{f}^{FCE}$. It should be noted that $\boldsymbol{f}^{FCE_h}$ is changed to $\boldsymbol{f}_6^{D}$ in sixth path. The progress can be denoted as (1),

$$\boldsymbol{f}_i^{FCE} = \begin{cases} \sigma(\delta(\boldsymbol{F}^{out}(Concat(\boldsymbol{f}_6^{D_h}, \boldsymbol{f}_i^{FPE}), \boldsymbol{W}_o))), & i = 5 \\ \sigma(\delta(\boldsymbol{F}^{out}(Concat(\boldsymbol{f}_i^{FCE_h}, \boldsymbol{f}_i^{FPE}), \boldsymbol{W}_o))), i = else \end{cases} \tag{1}$$

where, $\boldsymbol{f}_i^{FCE}$ denotes the i-th path FCE features. Concat$(\cdot)$ denotes concatenative operation, $\boldsymbol{F}^{out} \in \mathbb{R}^{W\times H\times C}$ refers to the output convolution, $\boldsymbol{W}_o$ denotes the parameters in convolution, $\delta(\cdot)$ denotes ReLU function and $\sigma(\cdot)$ denotes Sigmoid function. When $i = 1$, $\boldsymbol{f}_i^{FCE}$ denotes the edge features, which can be represented as $\boldsymbol{f}_{Edge}^{FCE}$ as well. To better refine the salient objects, as well. To better refine the salient objects, $\boldsymbol{f}_i^{FCE}$ is exported to compare with the salient truth, which is shown in Fig. 1.

TABLE I
DETAILS OF EACH CONVOLUTION WITH VARIOUS KERNEL

$\boldsymbol{F}^{EE}$	C_1			C_2			C_3		
1	1,3	0,1	256	3,1	1,0	256	3,3	1,1	256
2	3,5	1,2	512	5,3	2,1	512	5,5	2,2	512
3	3,5	1,2	512	5,3	2,1	512	5,5	2,2	512
4	3,7	1,3	512	7,3	3,1	512	7,7	3,3	512

b) Feature Progressive Extraction: As shown in Fig. 2, to extract the different layer features progressively, FPE adopts different dilation rates convolution in different layers, aiming to capture multi-scale multi-receptive-field context information. FPE follows a principle that shallow layers do not need to have large receptive field but the deep layers vice-versa. Previous researches usually adopts larger kernel size or higher number of convolutional layers to obtain larger receptive field but raise the parameters at the same time. To keep the parameters and enlarge receptive field, dilation convolution is adopted in FPE.

The larger the dilation rate is, the larger the receptive field is. Therefore, from $\boldsymbol{f}_1^{FPE}$ to $\boldsymbol{f}_5^{FPE}$, the dilation rates are set to 2/3/4, 2/4/6, 3/4/5, 3/4/6 and 3/5/7 respectively. 1×1 convolution means that only the number of channels is turned to one quarter, which keep part of the low-level features. The progressive setting of dilation rates means that the final extracted high-level features contain the features with scale and shape invariances from the low-level features.

C. Edge Enhancement Architecture with Various Kernels

Compared with the mainstay of salient map, the boundary features are sparse and less obvious. To overcome this issue, EEAVK is proposed to refine the edges of salient objects. Different from previous researches, various kernels in EEAVK is designed to obtain different scale edges, which aims to balance the scale between edge and mainstay of salient objects. In addition, different from the edge supervision in CFPM, EEAVK refines boundary features by specially designed network but not the supervision constraint, which improves the results further.

The four convolutional features of EEAVK are represented as $\boldsymbol{f}_i^{EE} \in \mathbb{R}^{W\times H\times C}$ As shown in Fig. 1, the progress of EEAVK can be represented as,

$$\boldsymbol{f}_i^{EE} = \boldsymbol{F}^{EE}((\boldsymbol{f}_{i+1}^{FCE} + \boldsymbol{f}_{Edge}^{FCE}), \boldsymbol{W}_E) \tag{2}$$

where, $\boldsymbol{F}^{EE}$ denotes the convolution with various kernel, and $\boldsymbol{W}_E$ is the parameters of $\boldsymbol{F}^{EE}$,

$$\boldsymbol{f}^{EEAVK} = \sigma(\delta(\boldsymbol{F}^{out}(\sum_{i=1}^{4} \boldsymbol{f}_i^{EE}, \boldsymbol{W}_o))) \tag{3}$$

where, $\boldsymbol{f}^{EEAVK}$ denotes the output features of EEAVK.

To better locate the boundary of salient objects, the various kernels is proposed in EEAVK, which can be denoted as $\boldsymbol{K}^{v'} \in \mathbb{K}^{l_1\times l_2\times C}$ and $\boldsymbol{K}^{v''} \in \mathbb{K}^{l_2\times l_1\times C}$ where $\boldsymbol{K}^{v}$ denotes the various kernel, l_1 and l_2 denote the length, $l_1 \in \{1, 3, 3, 3\}$ and $l_2 \in \{3, 5, 5, 7\}$. More specifically, the convolution with various kernels are shown in Table 1. In Table 1, $\boldsymbol{C}$ denotes three convolutional layers: C_1, C_2, C_3 and three followed ReLu layers. Each $\boldsymbol{C}$ includes the the kernel size, padding and channel number.

D. Loss Function

In this paper, loss function is divided into saliency loss function and edge loss function. Similar to other methods, cross-entropy loss is adopted to measure the similarity between the final saliency map and the ground truth, which is represented as follows,

$$\textbf{\textit{Loss}}^{Sal} = -\frac{1}{M \times N}\sum_{i=1}^{M}\sum_{j=1}^{N}(\boldsymbol{y}_{ij}\log\widehat{\boldsymbol{y}}_{ij}+(1-\boldsymbol{y}_{ij})\log(1-\widehat{\boldsymbol{y}}_{ij})) \tag{4}$$

where, $\boldsymbol{y}_{ij}$ denotes the ground truth in the location (i, j), $\widehat{\boldsymbol{y}}_{ij}$ denotes the final saliency map of SBN. Motivated by the significant applications of IoU boundary loss [32], this paper calculates the edge loss by IoU loss shown as follows,

$$\textbf{\textit{IoU Loss}} = 1 - \frac{2|\boldsymbol{C} \bigcap \widehat{\boldsymbol{C}}|}{|\boldsymbol{C}| + |\widehat{\boldsymbol{C}}|} \tag{5}$$

where, $\boldsymbol{C}$ and $\widehat{\boldsymbol{C}}$ denote the edge ground truth and final edge saliency map of SBN respectively.

IV. Experimental Results

A. Datasets

The performance of SBN is evaluated on five benchmark datasets: DUT-OMRON, DUTS, ECSSD, HKU-IS and PASCAL-S. DUT-OMRON contains 5168 high quality images, each of which has challenging complex background with one or more salient objects. DUTS contains 10553 images for training and 5019 images for testing. Following the previous work [6] [32], 10553 images in DUTS-Test is used for training in this paper. ECSSD contains 1000 meaningful and complex semantic images. HKU-IS has 4447 images with more than one disconnected salient objects. PASCAL-S has 5168 challenging images.

B. Evaluation Metric

Similar to most of the saliency detection methods, three standard metrics are used for evaluation in this paper, which include weighted F-measure (ωF_β) and mean absolute error (MAE) to evaluate SBN and other state-of-the-art methods. ωF_β is formulated as follows,

$$F_\beta = \frac{(1+\beta^2)Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{6}$$

where, precision and recall denote the ratio of salient pixels under different threholds between generated salient map and ground truth. β^2 is set to 0.3 as other mtethods do. MAE can be calculated by the following formula,

$$MAE = \frac{1}{M \times N}\sum_{i=1}^{M}\sum_{j=1}^{N}|\boldsymbol{y}_{ij} - \widehat{\boldsymbol{y}}_{ij}| \tag{7}$$

C. Implementation Details

SBN is trained on DUTS-Test dataset followed by [6] [32]. We do not use the validation dataset suggested by [6]. The weights of the rest part in SBN are initialized randomly. In the period of training, the learning rate is set to 5e-5 and reduced to one tenth for every 12 epochs. To reduce over-fitting during training, weight decay set to 0.0005 is also adopted in SBN. In addition, similar to other methods [6] [32], data augmentation techniques such as random rotating, random cropping, and random horizontal flipping are also adopted in this paper. The input image size is set to 256×256.

TABLE II
Metrics of Each the Added Module on ECSSD and DUTS-Test datasets

Module	**ECSSD**		**DUTS-Test**	
	ωF_β	MAE	ωF_β	MAE
ResNet	0.917	0.044	0.814	0.053
DetNet	0.925	0.041	0.815	0.053
DetNet+CFPM	0.936	0.039	0.835	0.042
DetNet+CFPM+EEAVK	0.944	0.038	0.863	0.040

It should be noted that the backbones are pre-trained in ImageNet.

D. Effectiveness of Each Module

To demonstrate the effectiveness of backbone, CFPM and EEAVK, we train backbone network, backbone with CFPM and backbone with CFPM and EEAVK respectively on ECSSD and DUTS-Test datasets in this section. As shown in Table 2, the result with each metric has a significant improvement. Compared with the backbone of ResNet, DetNet increases 0.9% and 6.8% of ωF_β and MAE on ECSSD dataset, and increases 0.1% of ωF_β on DUTS-Test dataset. Compared with DetNet without any extra architecture, the added CFPM in SBN increases 1.2% and 4.9% of ωF_β and MAE on ECSSD dataset, and increases 2.5% and 17.0% of ωF_β and MAE on DUTS-Test dataset. Compared with the added CFPM for DetNet, the added EEAVK in SBN increases 0.9% and 2.6% of ωF_β and MAE on ECSSD dataset, and increases 3.4% and 5.0% of ωF_β and MAE on DUTS-Test dataset. Results in Table 2 show the effectiveness of each module.

E. Comparison with State-of-the-art Methods

We quantitatively compared our SBN method with several state-of-the-art methods in recent three years: Pyramid Feature Attention Network (PFAN) [6], High-Resolution Salient Object Detection (HRSOD) [28], Output-guided Attention Module (OGNet) [24], Iterative and Cooperative Top-down and Bottom-up Inference Network (ICTBI) [19], Short Connections (DSS) [22], Embedding Attention and Residual Network (EARN) [29], Pyramid Attention and Salient Edges (PAGE) [4], Progressive Attention Guided Recurrent Network (PAGRN) [17], Recurrent Localization Network (RLN) [30], Reverse Attention Network (RAN) [25], Aggregating multi-level convolutional features (Amulet) [23].

a) Visual Comparison: As shown in Fig. 3, we compare SBN with other state-of-the-art methods. Obviously, the results of SBN outperforms other methods, which are closer to the ground truth in visual comparison. In detail, (1) SBN model detects small salient areas more accurate benefitting from the the application of DetNet instead of the traditional backbone such as ResNet or VGGNet (see Fig. 3 the 1, 3 rows). (2) With the help of feature coherence enhancement and feature progressive extraction in the connective feature pyramid module, SBN balances the different scale salient areas

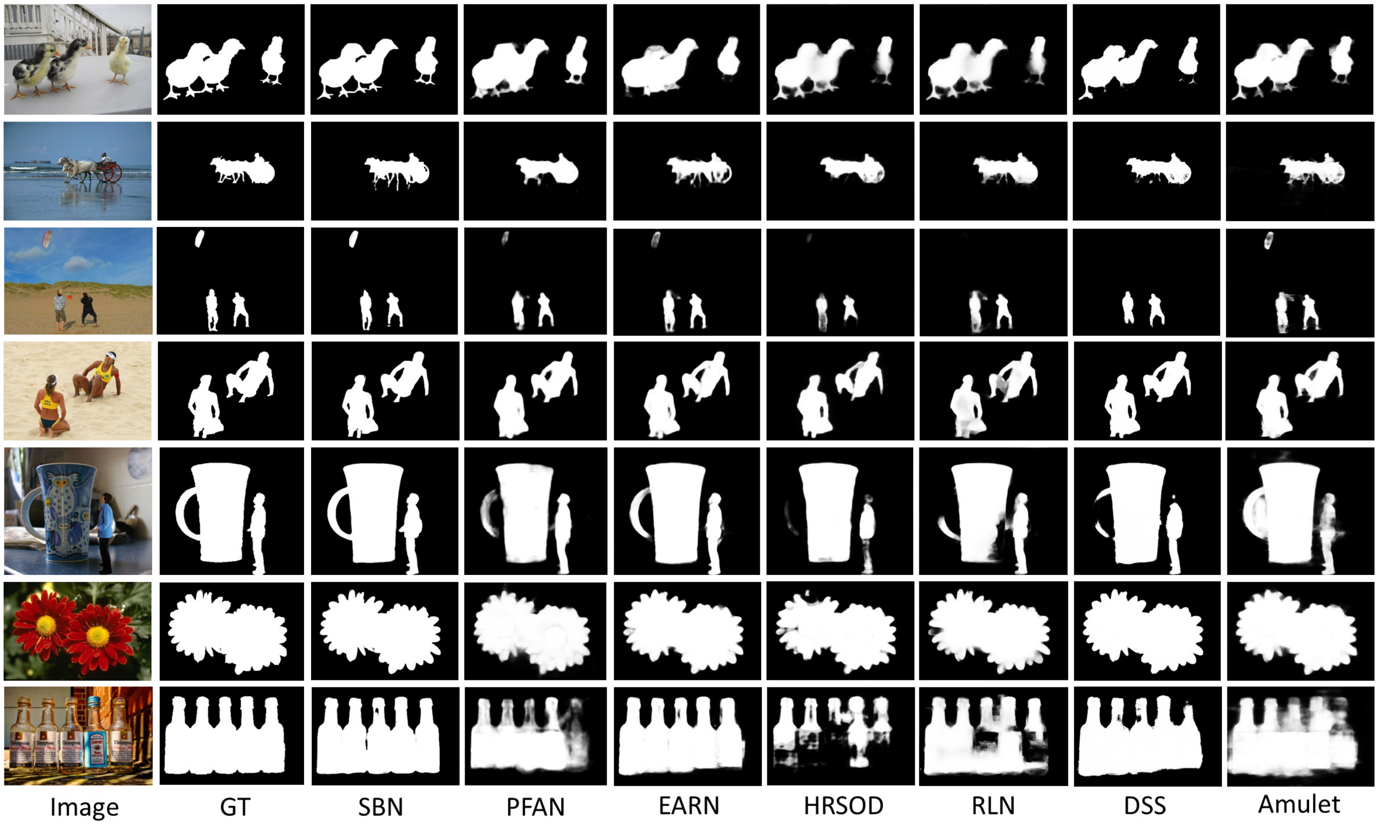

Fig. 3. Comparison with State-of-the-art Methods

TABLE III
COMPARISON WITH STATE-OF-THE-ART METHODS

Methods	ECSSD		HKU-IS		PASCAL-S		DUT-OMRON		DUTS-Test	
	ωF_β	MAE	ωF_β	MAE	ωF_β	MAE	ωF_β	MAE	ωF_β	MAE
Amulet [23]	0.868	0.058	0.854	0.052	0.763	0.098	0.647	0.098	0.737	0.085
PAGRN [17]	0.891	0.064	0.886	0.048	0.803	0.092	0.711	0.072	0.788	0.055
RAN [25]	0.918	0.059	0.913	0.045	0.834	0.104	0.786	0.062	-	-
DSS [22]	0.915	0.052	0.913	0.039	0.830	0.080	-	-	-	-
RLN [30]	0.903	0.045	0.882	0.037	-	-	0.709	0.063	0.768	0.051
OGNet [24]	0.916	0.047	0.916	0.041	-	-	0.743	0.066	0.807	0.047
HRSOD [28]	-	-	0.891	0.037	-	-	0.732	0.065	0.796	0.051
EARN [29]	0.921	0.057	0.916	0.040	0.845	0.095	0.802	0.061	0.844	0.059
PAGE [4]	0.924	0.042	0.918	0.037	0.835	0.078	0.770	0.066	0.815	0.051
ICTBI [19]	0.921	0.041	0.919	0.040	0.847	0.073	0.770	0.060	0.830	0.050
PFAN [6]	0.931	0.038	0.926	0.032	0.892	0.068	0.855	0.041	0.870	0.041
SBN	0.944	0.038	0.936	0.028	0.848	0.066	0.812	0.059	0.863	0.040

The best three results are shown in red, blue and green respectively.

better than other methods. For example, in Fig. 3 the 1, 3, 5 rows, SBN detects the small areas such as the feet of the ducks in the row 1 better than other methods. (3) By the aid of the edge enhancement architecture with various kernels, SBN locates the boundary information and extract the edge features more accurate than other methods (see Fig. 3 the 4, 6 rows). (4) Even though in the images with complex background, SBN generates the final salient maps more accurate and complete in form than other methods (see Fig. 3 the 1, 5, 7 rows). (5) Compared with other method, SBN highlights the salient object and suppresses the background regions better.

b) Quantitative Comparison: As shown in Table 3, SBN are compared with eleven state-of-the-art methods on five challenging datasets in terms of ωF_β and MAE. It can be seen that our SBN model wins on most datasets under the metrics and gets the best two results on all the datasets, which demonstrate our method is useful and effective. To be specific, test on ECSSD and HKU-IS datasets, SBN outperforms all of the other methods. On DUTS-Test dataset, SBN gets competitive results compared with PFAN but preforms much better than other methods. On PASCAL-S dataset, the ωF_β index of SBN is less than PFAN but the MAE is better, which means the confidence values of SBN is higher. However, though SBN gets a great improvement on very difficult and challenging DUT-OMRON dataset compared with other methods except PFAN, SBN still needs to be improved further on complex background images. In a word, according to the results in Table 3, the proposed model SBN is an effective and accurate network, which is able to detect salient objects well and make the network focus on the balance between different scale salient areas.

V. Conclusion

In this paper, a novel Scale Balance Network is proposed for locating large salient areas and recognizing small salient areas. In consideration of the over-large down-sampling factor in previous backbone networks, this paper adopts a specially designed backbone network for object detection called DetNet, which captures larger spatial resolution in deeper layers. Furthermore, a novel connective feature pyramid module is designed for balancing the scale between large salient areas and small salient areas, in which feature coherence enhancement improves the coherence between different convolutional layers and feature progressive extraction sufficiently leverages multi-scale and multi-level features. Besides, to refine the edge features, an edge enhancement architecture with various kernels is designed for locating better boundary features. Experimental results on each module added show the effectiveness of our proposed model. Our model outperforms other methods on ECSSD dataset and HKU-IS dataset, and achieves consistently superior performance in comparison with other state-of-the-art methods on other widely used datasets under different evaluation metrics.

References

[1] G. Lee, Y. W. Tai, and J. Kim. Deep saliency with encoded low level distance map and high level features. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pages 660668, 2016.

[2] Y. K. Hua, X. D. Gu. Group Loss: an efficient strategy for salient object detection. Communications in Computer and Information Science , 2019, 1142, 104111.

[3] L. C. Zhou, X. D. Gu. Embedding topological features into convolutional neural network salient object detection. Neural Networks. 2019

[4] W. Wang, S. Zhao, J. Shen , et al. Salient Object Detection With Pyramid Attention and Salient Edges. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019. pp: 1448-1457.

[5] N. Liu, J. Han, M.H. Yang. Picanet: Learning pixel-wise contextual attention for saliency detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018. pp: 30893098

[6] T. Zhao, X. Wu. Pyramid Feature Attention Network for Saliency Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019. pp: 3085-3094.

[7] K. M. He, X. Zhang, S. Q. Ren, J.Sun. Deep residual learning for image recognition. Proceedings of the IEEE Conference on Computer Vision, 2016. pp: 770778.

[8] S. Karen, Z. Andrew. Very deep convolutional networks for large-scale image recognition. International Conference on Learning Representations, 2015.

[9] Z. Tu, Y. Ma, C. Li, et al. Edge-guided Non-local Fully Convolutional Network for Salient Object Detection. arXiv preprint arXiv:1908.02460, 2019.

[10] J Su, J Li, Y Zhang, et al. Selectivity or invariance: Boundary-aware salient object detection. Proceedings of the IEEE International Conference on Computer Vision. 2019. pp:3799-3808.

[11] W. Wang, S.Zhao, J. Shen, et al. Salient Object Detection With Pyramid Attention and Salient Edges. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019. pp: 1448-1457.

[12] Z Li, C Peng, G Yu, X Zhang, Y Deng, J Sun. Detnet: A backbone network for object detection. European Conference on Computer Vision. 2018.

[13] O. Ronneberger, P. Fischer, T. Brox. Unet: Convolutional networks for biomedical image segmentation. In International Conference on Medical image computing and computer-assisted intervention, Springer, 2015. pages: 234241.

[14] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu. Global contrast based salient region detection. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015. pp:569582.

[15] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang. Saliency detection via graph-based manifold ranking. In Proceedings of the IEEE conference on computer vision and pattern recognition, 2013, pages 31663173.

[16] Z. Jiang and L. S. Davis. Submodular salient region detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pages: 20432050.

[17] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang. Progressive attention guided recurrent network for salient object detection. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pages: 714722.

[18] G. Li , Y. Yu. Visual saliency based on multiscale deep features. Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. pp:5455-5463.

[19] W. Wang, J. Shen, M. M. Cheng, et al. An Iterative and Cooperative Top-down and Bottom-up Inference Network for Salient Object Detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019. pp: 5968-5977.

[20] J Long, E Shelhamer, T Darrell. Fully convolutional networks for semantic segmentation. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015. pages: 34313440.

[21] Q Hou, M.M. Cheng, X Hu. Deeply supervised salient object detection with short connections. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. pp: 3203-3212.

[22] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr. Deeply supervised salient object detection with short connections. IEEE Trans. Pattern Anal. Mach. Intell. 2019.

[23] P Zhang, D Wang, H Lu, H Wang. Amulet: Aggregating multi-level convolutional features for salient object detection. Proceedings of the IEEE International Conference on Computer Vision. 2017. pp: 202-211.

[24] S Zhu, L Zhu. OGNet: Salient Object Detection with Output-guided Attention Module, IEEE Transactions on Circuits and Systems for Video Technology. 2019.

[25] S Chen, X Tan, B Wang, X Hu. Reverse attention for salient object detection. Proceedings of the European Conference on Computer Vision . 2018. pp: 234-250.

[26] H Wu, S Zheng, J Zhang. Fast end-to-end trainable guided filter. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. pp: 1838-1847.

[27] G Huang, Z Liu, L Van Der Maaten. Densely connected convolutional networks. Proceedings of the IEEE conference on computer vision and pattern recognition. 2017. pp: 4700-4708.

[28] Y Zeng, P Zhang, J Zhang, Z Lin. Towards High-Resolution Salient Object Detection. Proceedings of the IEEE International Conference on Computer Vision. 2019. pp: 7234-7243.

[29] S Chen, B Wang, X Tan, X Hu. Embedding Attention and Residual Network for Accurate Salient Object Detection. IEEE Transactions on Cybernetics. 2019.

[30] T Wang, L Zhang, S Wang, H Lu. Detect globally, refine locally: A novel approach to saliency detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2018. pp: 3127-3135.

[31] F Yu, V Koltun. Multi-scale context aggregation with dilated convolutions. International Conference on Learning Representations. 2016.

[32] Z Luo, A Mishra, A Achkar. Non-local deep features for salient object detection. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2017. pp: 6609-6617.