IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received September 28, 2019, accepted October 9, 2019, date of publication October 17, 2019, date of current version October 31, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2948062

Salient Object Detection: Integrate Salient Features in the Deep Learning Framework

QIXIN CHEN, TIE LIU, (Member, IEEE), YUANYUAN SHANG, ZHUHONG SHAO, AND HUI DING
College of Information Engineering, Capital Normal University, Beijing 100089, China

Corresponding author: Tie Liu (liutiel@163.com)

This work was supported in part by the National Natural Science Foundation of China under Grant 61603022, Grant 61876112, and Grant 61601311, and in part by the Support Project of High-Level Teachers in Beijing Municipal Universities in the Period of 13th Five-Year Plan under Grant CIT&TCD20170322.

ABSTRACT Salient object detection in complex environments brings the challenge from the collections of large number of training images for deep learning algorithm. It is difficult to collect the enough number of training data for varied salient objects in different scenes, and furthermore the salient objects are usually compared with the background. This paper proposes a novel method to integrate the salient features into the deep learning framework, and design a parallel multi-scale structure of the neural network to enhance the ability to detect salient objects. In addition, a multiple stage method is proposed to optimize the training process and make the datasets more prominent in the commonality of the salient features, which effectively reduces the difficulty of correctly distinguishing salient objects in complex environments. Experiments show that the proposed approach enhances the ability of neural networks to learn specified features and improves the detection effect of salient objects in complex scenes.

INDEX TERMS Salient object detection, salient features, deep learning, parallel multi-scale structure.

I. INTRODUCTION

In recent years, the visual saliency detection in images has become a topic of concern to scholars. Salient object detection aims to locate the most attractive visual target from images. The relevant research on salient object detection has been applied to a variety of practical scenarios [1], [2]. Moreover, visual attention has been widely leveraged as a pre-processing step in various computer vision fields, such as object detection [3], [4], object segmentation [5], [6], video summaries [7] and video object rebuilding [8], while salient object detection method provides effective promotion.

The human visual system is usually simulated to determine the most important and significant objects in the image, while salient object detection method allows to instantly focus on the most interested in the whole image. The representative saliency computational model is first proposed by Itti *et al.* [9], and the salient object detection is then proposed and studied by Liu *et al.* [15] which inspires the research on salient object detection in the following years. However, accurate salient object detection in complex and diverse environments as shown in Figure 1 still face the challenges, such as incomplete detection, confusion of targets and backgrounds, and misjudgment of the saliency map.

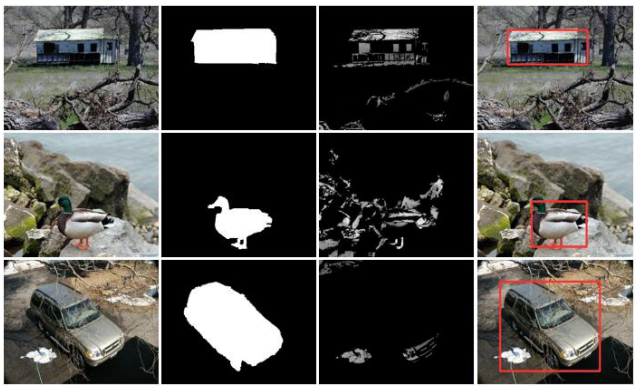


FIGURE 1. Salient object detection in complex environments. From left to right: Original images, ground truth maps, the defective effect results, and the results of our approach.

Saliency-related detections approaches are divided into top-down method and bottom-up method [10]. The top-down methods are based on human experiences of visual saliency, and the bottom-up methods are based on the capability of extracting salient features in tagged images. It is difficult to

The associate editor coordinating the review of this manuscript and approving it for publication was Byung-Gyu Kim.

separate these two processes completely because any single one method faces the challenges to distinguish salient objects in a complex environment.

Due to the fact that the traditional approaches with low-level manually extracted features are difficult to master the deep semantic information of images, they have been replaced by more and more deep learning approaches in recent years. At the same time, the accuracy of salient object detection (SOD) has also been greatly improved. However, it cannot be ignored that SOD is more special than other object detection tasks. The salient objects have strong subjectivity on assessing and lack completely objective features to learn for the computer. In the existing deep learning approaches, in order to obtain more complete graphics and semantic information in images, the network structures used by scholars are more and more complicated. End-to-end learning greatly enhances the autonomy of the neural network to acquire feature maps, enabling them to learn effective information in images. Due to the strong ability of concluding features of the neural network, manually extracted features only occasionally appear in the problems as lack of data or limited computational resource. But at the same time, the flaws which are difficult to control have become a barrier between the deep learning and humans. Excessive deep features are not helpful for detecting salient objects in complex environments. Not all features are equally important in saliency detection, and even some unrelated features can cause interference [11]. Even if researchers can obtain the misleading information learned by neural networks through the method of interpretability [30], [31] or others, it is still difficult to control the neural networks to learn the verified effective salient features.

Therefore, this paper proposes a method and designs a parallel multi-scale structure that enables the selected effective features to play a more important role in the end-to-end neural network. While maintaining the main network training original input images, the branch network is added to extract selected salient feature maps and train them. In order to break the learning autonomy of neural network, we enhance the influence of selected features by concatenating the feature map with saliency information into the main network. We repeat this operation in the layers of different depths. The parallel training structure also keeps the original information of the image intact. The structure we designed allows the selected features to affect the deep layers of the neural network and solves the problem that the known useful features are difficult to make effective impact to the outputs. At the same time, the multi-scale structure can combine shallow graphic features with deep semantic information to make the predict results closer to human judgment.

Visual saliency is highly influenced by the environment, so the evaluation index for it is significantly different from other target detections. For example, when facing the scene with small objects, multiple targets, blurred objects, occlusion objects or other complex situations, there are different judgments for salient objects. The same objects in different environments usually show different degrees of saliency. Any object is salient or not should be judged according to the overall environment. But in any case, the only principle for assessing the saliency of objects should always be whether it is visually salient to the eye. In this regard, we have researched a large number of saliency data and designed a stepped detection method specifically for detecting salient objects. It effectively improves the ability of computers to distinguish salient objects in complex scenes.

All in all, our contributions are as follows: First, we summarize the factors which are more important in assessing the saliency and extract the corresponding feature map. Secondly, we construct a parallel multi-scale structure that allows the influence of selected features in the neural network can be artificially enhanced. Third, we propose a stepped detection method specifically for salient objects. In addition, we propose a method of optimizing the training set so that the neural network can focus on training more salient regions.

II. RELATED WORK

In this section, we describe the irreplaceability of salient features of SOD in three ways. First, we analyzed the characteristics of salient objects. Second, we point out the difficulties of distinguishing salient objects. Finally, we discuss the advantages and disadvantages of the current saliency detection approaches.

A. CHARACTERISTICS OF SALIENT OBJECTS

The purpose of SOD is using machine learning to enable the computer to focus on the targets that are most interesting to human vision in the image. These targets usually include humans, animals, cars, text or others which can visually stimulate to our eyes. They usually have similar features that enable human vision to consciously focus on them in a very short time.

At present, the approaches based on convolutional neural network become popular for more and more researchers, that fewer and fewer pay attention to the characteristics of saliency itself. However, regions with more obvious salient features will be first focused and analyzed by humans. Through in-depth researching and testing to the classical non deep learning saliency detection approaches such as Itti *et al.* [9], HS [12], DRFI [13], wCtr [14] and Liu *et al.* [15], this paper concludes that salient objects are usually significantly different from non-salient objects in three aspects: bright differences, color distribution and layout structure in the image. Features that play a more critical role will vary in different scenarios. However, making salient features play a more prominent role in training and detecting process will inevitably increase the accuracy of detection.

B. DISCRIMINATION OF SALIENT OBJECTS

In machine learning, people give computer training data and the ground truth, then computers can perform some kinds of detection tasks. However, salient object detection has strong subjectivity that even humans cannot guarantee a unified

judgment on the salient targets of the image, not to mention the machine. Therefore, providing more reliable salient regions to the computer is more conducive to learn salient features rather than the characteristics of the object itself. Scholars label the ground truth map GT of salient objects through the binary map:

$$GT(x, y) = \begin{cases} 1, & \text{if } I(x, y) \in \text{Salient region} \\ 0, & \text{if } I(x, y) \in \text{Non-salient region} \end{cases}, \quad (1)$$

where $I(x, y)$ is the corresponding pixel in the original image. But in fact, each part of the object shows different degrees of saliency. The lack of saliency parts often brings many irrelevant factors. In the binarized ground truth map, the overly detailed object regions make the training network bias toward the characteristics of the object itself, interfering the judgment of the salient features. Salient regions are strongly influenced by the environment that small changes in brightness or color may cause large fluctuations to their saliency level. Scholars should be more cautious in assessing the saliency of each object.

C. NETWORK STRUCTURE IN SALIENCY DETECTION

The existing approaches of saliency detection have continuously made breakthroughs in neural networks, drawing excellent inspiration from various fields and achieving better and better results. In recent excellent papers, Li *et al.* [16] and Zhang *et al.* [11] used a double-sided network structure in their saliency detection approaches. They use deep feature maps to feedback back to generate predicted values, which improves the reading ability of semantic information. But many SOD approaches, including them, ignore the diversity and subjectivity of salient targets. Zhang *et al.* [32] integrated deep saliency and unsupervised saliency as input, diversifying the characteristics learned by the network. Zhang *et al.* [26] added reformulated dropout structures into the network to enhance the uncertainty of feature extraction. Their approaches take into account the uncertainty of salient features and acquire richer salient features.

There are always outstanding scholars to propose methods for optimizing SOD network, but the detection accuracy has entered a bottleneck range for a long time. The existing deep learning methods lack bias to the specified features, so that it is difficult for humans to guide the learning process of neural networks through owned experience. Recently, Zhao *et al.* [33] and Su *et al.* [34] utilized the difference between the feature maps of different depths to detect the edges and interior of the salient objects separately, and obtained a clear detection result. The success of these approaches demonstrates that the suggestion of extracting features is effective to neural networks in eliminating negative interference. Then, it is feasible and necessary to design a network structure that allows selected features to play a more prominent role in deep learning. Finding the salient features with commonality rather than the features of the object itself, can enhance the ability of the SOD network to adapt to complex environments.

In addition, the design of multi-scale network structures is increasingly used in saliency detection. Liu and Han [17] proposed a two-stage deep network, which uses shallow information to refine the coarse prediction maps in the deep layers. Zhang *et al.* [18] used a loop module to combine five different scale convolution features to generate a prediction map. The multi-scale structure has sufficient understanding of the information in different depths. Many existing experimental results also confirm the validity of multi-scale structure which has become a necessary method in SOD.

III. OUR APPROACH

In this chapter, we present a method aiming specially for detecting salient objects which with strong subjectivity. Our approach can integrate the known salient features into neural networks. It helps to solve the problem that neural networks are difficult to find salient features correctly and locate salient regions in complex environments. The targeted design of this paper is mainly implemented in the following three parts: In Section A, we extract feature maps which are closely related to the saliency of images. In Section B, we introduce our parallel multi-scale network structure and the design ideas in detail. In Section C, we have designed a stepped detection module specifically to identify salient targets.

A. FEATURE EXTRACTION

This paper argues that finding the factors most relevant to saliency can effectively enhance the ability of neural networks for learning salient features. As mentioned in Section II.A, salient objects usually show obvious differences from non-salient objects in bright differences, color distribution and layout structure. Therefore, we extracted the feature maps $\{f_L, f_V, f_E\}$ of three different factors from the original image as shown in Figure 2. Since the extracted feature maps will also pass through the convolutional neural network, they do not have to be processed complexly, thus reducing the time spent.

1) LUMINOSITY MAP

Luminosity is widely recognized as an image attribute that most intuitively affects visual attention, and a luminosity map can fully present a bright distribution of an image. In the color space Lab, the luminosity component (channel L) is completely separated from the color components (channel a, b), reducing mutual interference between different factors. As shown in Figure 2(b), in terms of the visual impact, the luminosity map (f_L) is the closest to the human eye, and it can accurately quantify the brightness of each pixel.

2) COLOR FREQUENCY FEATURE MAP

In the image, the frequency and density of the color is the most intuitive color distribution information that the human eye receives. In order to get a more relevant color feature map, we convert the original image to the color space HSV. Where

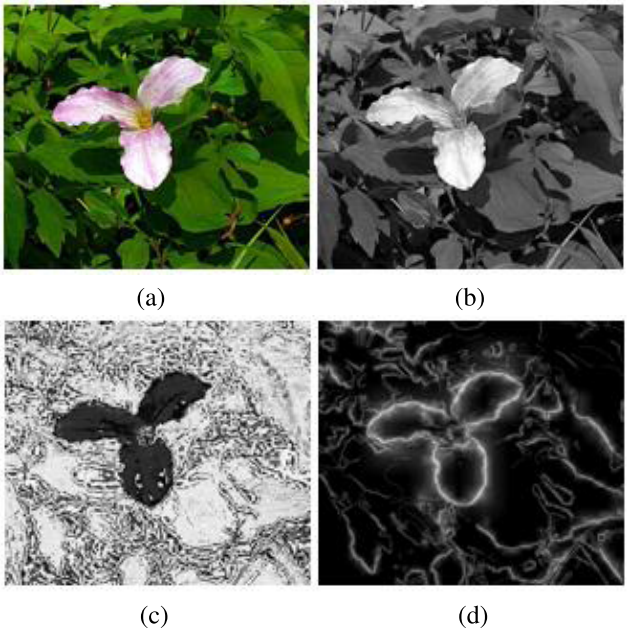


FIGURE 2. Extract three different sets of feature maps from the image. (a) Original image. (b) Luminosity map. (c) Color frequency feature map. (d) Multi-scale edge map.

the hue component (channel H) expresses the overall color tendency of the image. The saturation component (channel S) has an influence on the vividness of the color. And the value component (channel V) controls the brightness of the image, which is similar to the luminosity component. We quantize the image according to the influence on the color change, getting a new feature map f_Q:

$$f_Q(x,y) = H'(x,y) \times N_S \times N_V + S'(x,y) \times N_V + V'(x,y), \quad (2)$$

where H', S' and V' are the quantized values of the H, S and V channels, and N_S and N_V are the number of stages quantized by S and V channels. At this time, the feature map f_Q has a strong monotonic correlation. We weight the statistics to get the frequency feature map f_V:

$$f_V(x,y) = \sum_S n_S(x,y) \times k_S, \quad (3)$$

where $n_S(x,y)$ is the amount of pixels in the feature map f_Q which have a distance S from the point (x,y) and have similar color, and k_S ($k_S \in [0.01, 1]$) is the weight parameter when the distance is S. At this time, the feature map f_V no longer appears as a specific color, but instead presents the color distribution of overall the picture, as shown in Figure 2(c). The color frequency feature map usually shows the characteristics of strong regional. The brightness difference of this map can greatly help to distinguish different salient regions.

3) MULTI-SCALE EDGE MAP

The edge feature is an effective feature that can make us grasp the spatial position relationship of each object in the image. Edge feature is not sensitive to small changes in luminosity or color, so that it is useful for position determination and region segmentation of objects. The most popular method "super-pixel segmentation" has the problem of too long running time that it lacks competitiveness in the practical application of salient object detection. After transforming the original image into six scales of {1, 1/2, 1/4, 1/8, 1/16, 1/32}, we use the Canny operator to calculate the edge feature map separately. Then resize them to the size of the inputs and add them together to get the edge feature map f_E:

$$f_E = \frac{1}{N} \sum_{n=1}^{N} Resize(Edge_C(f_n)), \quad (4)$$

where $Edge_C(\cdot)$ represents the edge feature map calculated by the Canny operator, f_n is the image of the n^{th} scale and N is the amount of different scales. As shown in Figure 2(d), the graphical description capability of the edge feature map after multi-scale operation is greatly enhanced. In addition, it also has less loss in runtime.

The above feature maps capture the salient features of the image in completely different ways. Tested in known saliency datasets, these feature maps show significant differences between the salient and non-salient regions, indicating that they have the ability to distinguish salient features.

B. PARALLEL MULTI-SCALE STRUCTURE

Aiming at the problem of making the manually extracted feature maps fully join into the neural network, this paper designs a parallel multi-scale structure. Our structure solves the problem that specified features cannot be given enough attention in the neural network. It enables critical features to make more influence, enhancing the ability of the neural network to learn saliency information of images.

The main framework of our approach is shown in Figure 3. Our basic structure is quoted from the excellent well-recognized target detection approach YOLOv3 [19], which is an object detection approach that is known for fast speed. The main network for downsampling input images uses the front five residual convolution blocks of Darknet-53, including 23 residual units, totally 52 convolutional layers. In such a structure, the downsampling operation composed of convolutional layers instead of pooling layers retains more details. And the residual module ensures that the network depth is sufficient to extract valid features and avoids gradient disappearance or gradient explosion.

In order to integrate the extracted feature map into the neural network and play the effect, we concatenate the extracted feature maps and integrated them into a salient feature map f_I', which has the same size as the original image through convolutional layers:

$$f_I' = Conv\{f_L, f_V, f_E\}. \quad (5)$$

The branch network used to train the extracted feature map f_I' has the same structure as the main network. This design

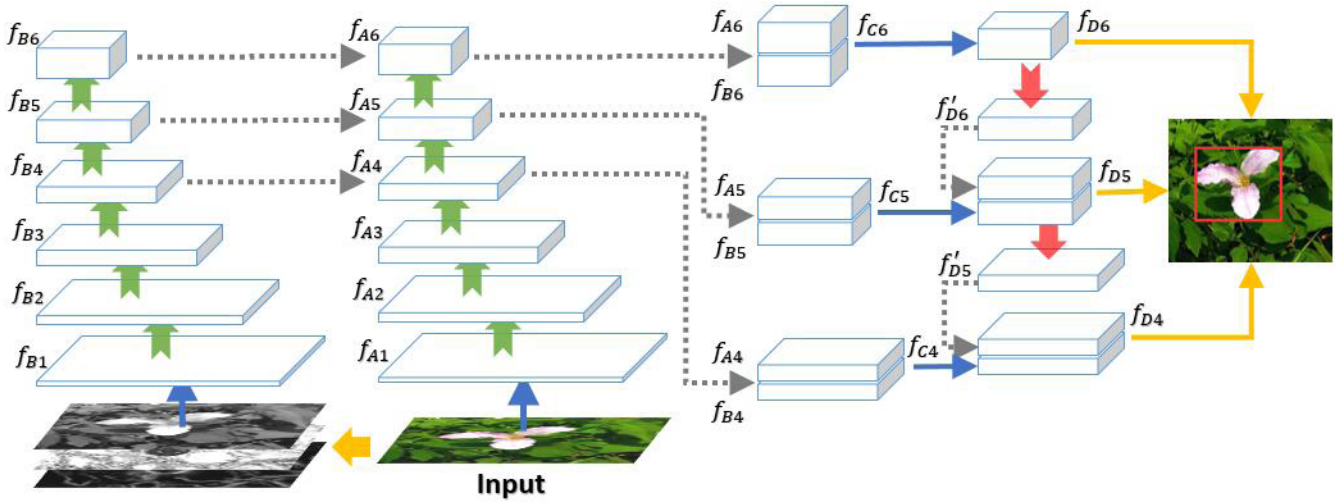


FIGURE 3. The main framework of our approach. Some structures such as convolutional layers and residual structures are omitted.

enables feature maps of different depths to find branch network maps of the same size. As shown on the left side of Figure 3, these two streams form a group of parallel networks ($f_{A1} - f_{A6}$ and $f_{B1} - f_{B6}$). They give us the ability to make directed guidance to the feature maps which are trained to any stage. We concatenate the two sets of feature maps in the same stage in the deeper layers and integrate them through the convolutional layers:

$$f_{Cn} = Conv\{f_{An}, f_{Bn}\}, \quad n \in \{4, 5, 6\}. \tag{6}$$

But in the shallow layers (layer 1 to 3) of the network, it is necessary to maintain their independence of training to ensure salient features extracted accurately. Multiple integrations can enhance the impact of salient features at different stages, fully guiding the training process of the original image. At the same time, the feature maps reaching the deep layers makes the important selected features have a more direct impact to the output and play a greater role.

In continuous downsampling operations, the resolution ratio of the image is getting lower and lower, and the detail message is getting more and more blurred. The shallow feature map in the convolutional neural network has more detailed information, and the edge features are relatively clear and complete. The deep feature map contains the semantic information that the computer interprets. Both them are critical for detecting salient objects and complement each other. We carry out the upsampling operation to the deepest feature map f_{D6}, and generate a higher resolution feature map f'_{D6} with clearer semantic information. Concatenating f'_{D6} with f_{C5} of the same scale in the shallower layer, the generated feature map f_{D5} has richer information for predicting salient objects. Continuous feedback from deep layer information and associated with features of shallow layers enhances the ability of high resolution feature maps to learn semantic information, and effectively improves the accuracy of the prediction.

C. STEPPED DETECTION MODULE

After carrying out experiments and analysis on a large number of saliency data, this paper proposes a more effective detection method for the characteristics that the saliency is highly susceptible to environmental influences. Salient objects lack objective fixed features so that the characteristics of a single target are not sufficient to distinguish whether the object is sufficiently salient in the complex scene. Salient object detection should be regarded as the task of the whole image. By analyzing the salient level of each region in the image, the most interesting content of the eye can be found through judging the relationship between the whole and the part.

In the detection section, our approach also learned the method of predicting Bounding Boxes from YOLOv3. We get the 9×9 anchor points with average distribution in the image and predict six parameters $\{S, X, Y, W, H, C\}$ for each predicted box centered on each anchor point. Where S indicates the judgment of whether the predicted box contains salient objects, X and Y indicate the coordinate of the center point of the predicted box, W and H indicate the width and height of the predicted box, and C indicates the confidence of the neural network to determine the predicted box:

$$C = S \times IOU_{Predction}^{Groundtruth}, \quad S \in \{0, 1\}, \tag{7}$$

where $IOU_{Predction}^{Groundtruth}$ is used to indicate the coincidence between the reference and the predicted box [20]. We predict the integrated feature maps at the deeper three scales $\{f_{D6}, f_{D5}, f_{D4}\}$, and obtain a large number of predicted boxes and corresponding predicted confidence at different scales. We merge the predicted regions with mostly overlapped and preserve those with higher confidence. For the predicted boxes at the same position of different scales, we use the weighting method to bias the coordinates of the predicted box to the shallow feature map and bias the predicted confidence to the deep feature map. It is found from statistical data that

FIGURE 4. The detection results of the objects with different saliency level. The red represents strongly salient objects, the blue represents weakly salient objects, and other parts belong non-salient regions.

the final predicted confidence for each region is positively correlated with the saliency level.

To assess the saliency level of the objects, we propose a stepped detection module. The predicted confidence is used as a reference to smoothly set two thresholds, which classify the objects of different salient degree into three categories: strongly salient objects, weakly salient objects and non-salient regions. In Figure 4, we used the red boxes and blue boxes to mark the detected strongly or weakly salient targets. Among them, the strongly salient targets are always focused by the eye in any images. They have a strong visual impact and are usually more familiar to humans. The weakly salient target also has valuable information, but it is often overlooked because of lack of visual impact when it is in the same environment with a stronger target. People will always collect the secondary information after finishing the most important parts. But, in the scenario of Figure 4(e), when there is no strongly salient object in the image, those weaker ones will contain most of the valid information of the image. Even with a lack of visual stimulation, people can still instantly focus on the targets they are most interested in. All regions below the minimum threshold are considered to be difficult to attract attention to the human eye that we define them as non-salient regions. These regions usually have the characteristics of lack of focus, dull in colors, blurred pixels or close to edges. It is difficult for people to stay focused on such a position.

Because the stepped detection module has the ability to assess the image as a whole, our approach can also accurately separate the images that do not have salient objects from datasets, as shown in Figure 4(g).

IV. EXPERIMENTAL RESULTS

The approach proposed in this paper aims to detect salient objects in the image and mark them. In recent years, in order to enrich the diversity and practicability of SOD, many scholars have extended the research to various fields. Focusing on the definition of the saliency, the extended slender branches of the salient objects and other isolated parts as shown in Figure 5 are hardly considered to be visually salient, even though they are still part of the salient object. Each part of a salient object will shows different degree of saliency in the image. In the end, the experimental results confirm our view that removing those redundant branches and fragments can get more accurate detection results.

FIGURE 5. Remark the salient regions with redundant branches. (a) Original image. (b) Ground truth map. (c) Relabeled map.

A. EXPERIMENTAL SETUP

Training parameters: Our experiments were performed on a server with a Tesla K40C GPU. We set the total of iterations to 15000, and the learning rate is 0.001 with twice decay.

Training datasets: We summarize the reliability of labeled ground truth maps and the coverage of the image content of widely used saliency datasets. Then, we randomly selected 2500 images in MSRA-B [15] (5000 in total) and 2500 images in HKU-IS [21] (4447 in total), totaling 5000 images and their ground truth maps for the training set and the validation set. The remained datasets are used as test sets. As shown in Figure 5, for the ground truth maps of the training set, the rectangular marker box is used for relabeling. We cut off the slender branches and isolated fragments which protrude from salient objects, only retaining the more salient parts as the ground truth regions for training. This allows the neural network to focus more on the more salient parts rather than the characteristics of the object itself.

Test datasets: Our approach is tested on the widely used saliency datasets, such as MSRA-B, HKU-IS, EC-SSD [12], PASCAL-S [22] and DUT-OMRON [23]. These datasets are rich in content and contain a large number of images with complex environments. Therefore, the experimental results are representative.

Evaluation index: In saliency detection, F-measure (F_β) and mean absolute error (MAE) are commonly used as indicators for evaluation. These two indicators provide a comprehensive assessment of the accuracy for salient object detection. F-Measure is defined as:

$$F_\beta = \frac{(1+\beta^2) \times precision \times recall}{\beta^2 \times pricision + recall}, \tag{8}$$

where $\beta^2 = 0.3$. MAE is defined as the average pixel-wise absolute difference between the ground truth map and the

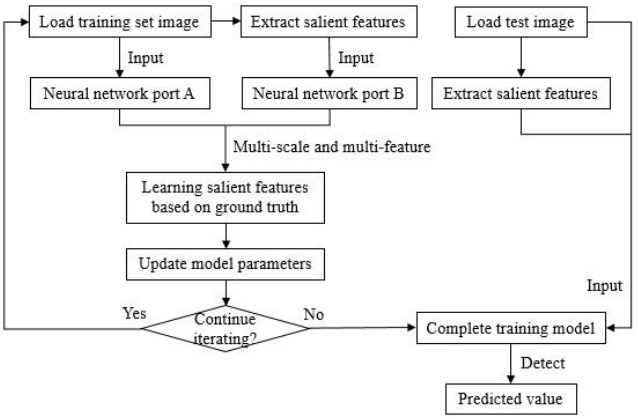


FIGURE 6. The flow chart of the experiment.

TABLE 1. Test results of our approach and other advanced approaches on saliency data sets.

	HKU-IS		EC-SSD	
	F_β	MAE	F_β	MAE
DRFI [13]	0.776	0.167	0.782	0.170
RFCN [24]	0.896	0.073	0.899	0.091
DHSNet [17]	0.892	0.052	0.907	0.059
DCL+ [25]	0.904	0.049	0.901	0.068
UCF [26]	0.823	0.061	0.844	0.069
Amulet [18]	0.843	0.050	0.868	0.059
DSS+ [27]	0.913	0.039	0.915	0.052
MSRNet [16]	0.916	0.039	0.913	0.054
Zhang's [11]	0.886	0.048	0.891	0.064
ASNet [28]	0.920	0.035	0.928	0.043
Chen's [29]	0.913	0.045	0.918	0.059
Ours	0.922	0.034	0.929	0.043

prediction map:

$$MAE = \frac{1}{H \times W} \sum_{x=1}^{H} \sum_{y=1}^{W} |S(x, y) - G(x, y)|, \qquad (9)$$

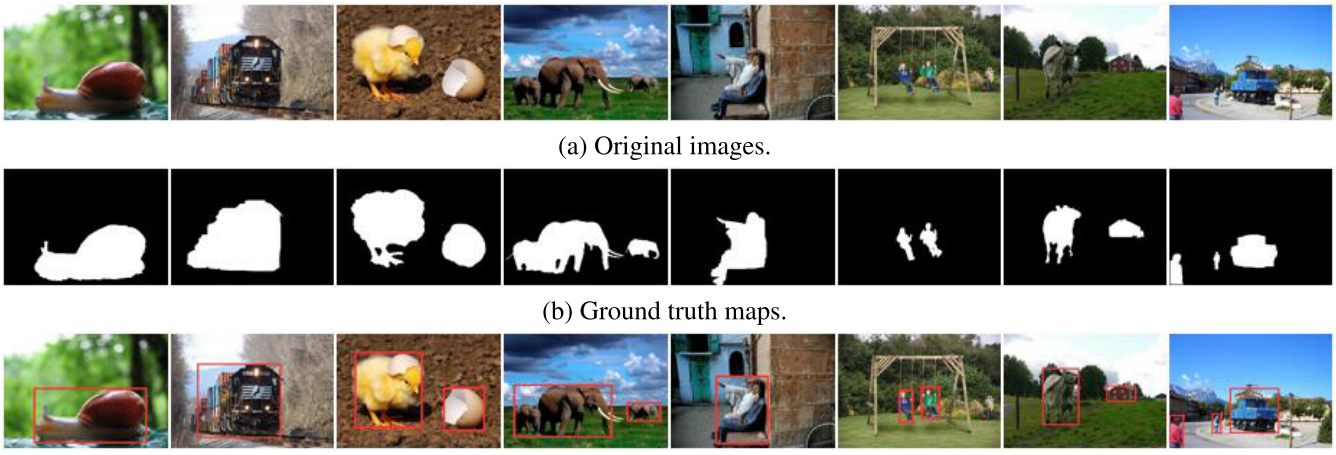
where W and H are the width and height of the image.

B. DETECTION RESULTS

The experimental flow chart is shown in Figure 6, including the two processes of training the model and detecting the images. The randomly selected 5000 training set images and the corresponding re-labeled ground truth data are input into our network. During more than 100 hours, the training process results in a loss stably less than 0.1 and a trained model for salient object detection.

Our approach performs an average detection time of 0.12s per image, having advantages in the speed compared to other SOD approaches with the same precision. Ours has higher practical value. Figure 7 shows the results of our approach on saliency datasets. Our approach is able to quickly and accurately find salient objects in an image. Even in the face of difficult targets such as small objects, multiple targets, blurred objects, and occluded objects, ours can avoid errors or missed detections. Moreover, our approach can accurately evaluate the saliency relationship of each region of the whole image. It effectively solves the problem of being difficult to distinguish salient targets in complex environments.

Our approach is compared with the best saliency detection approaches from top-level conferences on frequently-used saliency datasets. In addition to DRFI is a classic traditional approach, the others are all excellent deep learning approaches in recent years. As shown in Table 1, with the same size of training sets, our approach is able to achieve better results in two commonly used saliency data sets, HKU-IS and EC-SSD. In these two datasets, the images are richer in content and have backgrounds with varying complexity. Especially, these two datasets have more reliable ground truth labels. Even in the scenes with high detection difficulty as shown in Figure 8, the marked salient regions are still less controversial. In such a situation with low contrast or severe interference of the same type, our approach still accurately



(a) Original images.

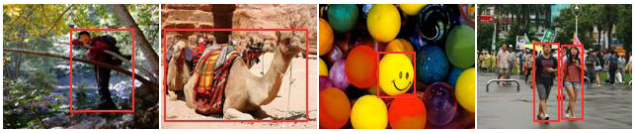
(b) Ground truth maps.

(c) Our results.

FIGURE 7. Our detection results on the saliency data sets.

(a) Original images.

(b) Ground truth maps.



(c) Our results.

FIGURE 8. The images with high detection difficulty in datasets HKU-IS and EC-SSD, and our detection results.

detects salient objects. Tested in other datasets, our approach is also close to the best current test results. Compared to the mentioned high-level SOD approach ASNet [28], we use known effective features to replace the role of their fixation maps which are more difficult to collect. Moreover, our approach is more intuitive that it can target to solve the problem with known defects. It is worth mentioning that, in the case that the pixels in the detection boxes are 59% more than pixel-level methods, the MAE values of our approach are still close to others, showing extremely high reliability.

C. EFFECTIVENESS OF OUR APPROACH

The proposed approach aims to enhance the network's ability of detecting salient objects by integrating selected salient features into the neural network. The characteristics of saliency are our main research points. The designed network structure and other modules all serve to enable the neural network to learn salient features more effectively. As can be seen from Figure 9, our approach achieves a significant improvement in each data set compared to the approach YOLOv3 which has the same basic structure as ours.

Especially, the changes are more obvious in the images with complex environments. As shown in Figure 10, some defective detections have occurred in the results of YOLOv3. The proposed parallel multi-scale structure solves these problems by obtaining more accurate saliency information. Some detection results which close to the ground truth are exhibited in Figure 10. In the first and second images, it is proved that our approach can accurately analyze the saliency level of each object and avoid the missed detection. In the third image, our approach extracts more accurate salient features in low contrast environments, eliminating the targets that cause strong interference. In the fourth image, more effective training data allows us to capture key regions of salient objects, avoiding the inaccurate detection caused by complex environments. The effectivity of our design supports that it is possible for computer to understand more human experience. So that our

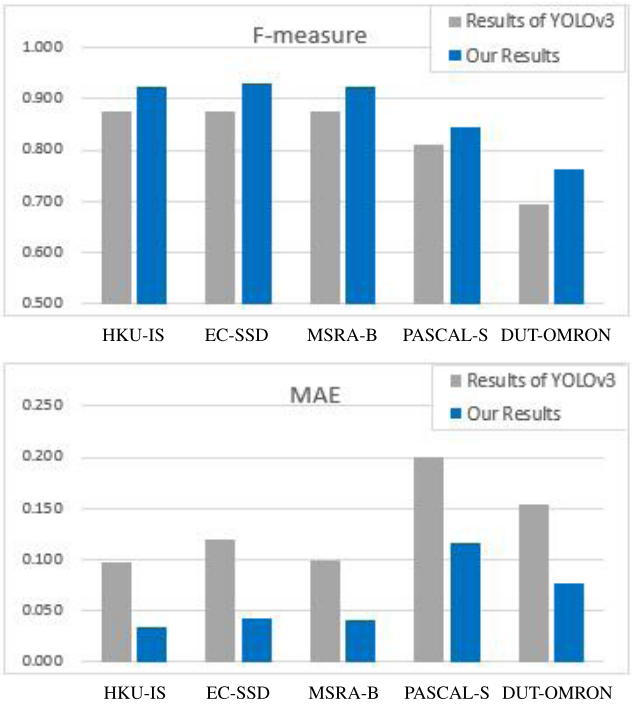
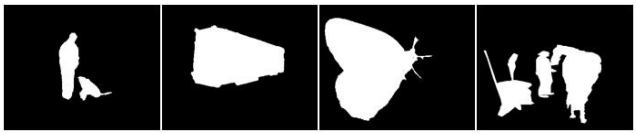


FIGURE 9. Our approach is compared F-measure and MAE with YOLOv3.

(a) Original images.



(b) Ground truth maps.

(c) The results of YOLOv3

(d) Our results.

FIGURE 10. The detection results in complex environments of our approach and YOLOv3.

approach is able to be extended to other automation works which still require human assistance, such as behavior prediction and self-driving.

D. ABLATION ANALYSIS

We have validated the effectiveness of each module proposed in this paper. As shown in Table 2 and Figure 10, the gradual

TABLE 2. The validation of different modules of our approach.

Saliency data sets	Basic structure from YOLOv3		Parallel multi-scale structure		Parallel multi-scale structure + Stepped detection module		Our final results	
	F_β	MAE	F_β	MAE	F_β	MAE	F_β	MAE
HKU-IS	0.876	0.098	0.888	0.054	0.892	0.051	0.922	0.034
EC-SSD	0.874	0.119	0.897	0.061	0.902	0.058	0.929	0.043
MSRA-B	0.877	0.099	0.890	0.063	0.896	0.058	0.925	0.041
PASCAL-S	0.811	0.200	0.833	0.116	0.839	0.115	0.843	0.116
DUT-OMRON	0.695	0.154	0.720	0.082	0.745	0.080	0.764	0.077

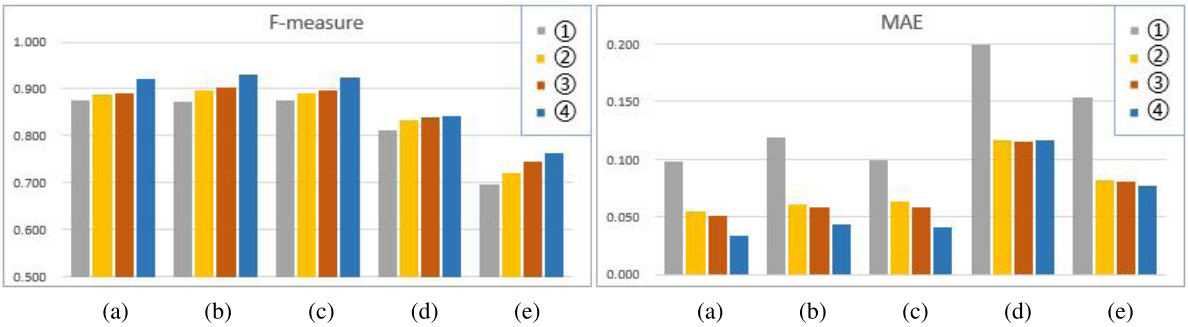


FIGURE 11. The influence of different modules in our approach on F-measure and MAE. (a) HKU-IS (b) EC-SSD (c) MSRA-B (d) PASCAL-S (e) DUT-OMRON ① Basic structure from YOLOv3. ② Parallel multi-scale structure. ③ Parallel multi-scale structure + Stepped detection module. ④ Our final results.

improvement of the experimental results confirms that each of our designs is reasonable and effective.

1) PARALLEL MULTI-SCALE STRUCTURE WITH SALIENT FEATURE MAPS

First, we verified that our parallel multi-scale structure can contribute to integrate selected features into neural networks. In this part, we add the parallel multi-scale structure into YOLOv3 and compare with the original YOLOv3. Our structure is not very complex, but it is characterized by the ability of carrying the required features into each layer of the neural network. Compared to other approaches, especially those ignoring salient features, integrating effective features gives ours a comprehensive advantage. The first two sets of data in Table 2 shows that the designed structure can increase the F-Measure by about two percent in each dataset. This proves that in deep learning approaches of SOD, there is indeed a problem that it is difficult to extract salient features effectively.

2) STEPPED DETECTION MODULE

The stepped detection module is specifically designed for the difference between SOD and other detection tasks. As can be seen from the third set of data, which indicate the results after adding the stepped detection module, the new module gives each prediction region the ability to evaluate its saliency level in the whole image. This allows the detection network to have a more accurate judgment in complex environments.

3) RELABEL THE TRAINING SET

Using the training set labeled with more attention to saliency, which removes slender branches and isolated fragments of salient objects, can effectively switch the network's focus to salient features. At the expense of less recall, the precision of detection is significantly increased. It is not difficult to see from the last set of data that the total detection ability of salient objects has been greatly improved. For our approach focuses on researching salient features, it is more important than others to locate an effective region.

V. CONCLUSION

In this paper, we explore the characteristics of salient objects and designed a network structure which can integrate selected features into the learning algorithm. The proposed method helps to solve the problem that convolutional neural networks are difficult to accurately extract salient features in complex environments. Compared to other SOD approaches, known effective features in ours are able to play a more critical role in neural networks. Moreover, the combination of multi-scale and global assessment enables the algorithm to be similar to the criterion of human visual attention mechanism in salient object detection. The experimental results indicate that our approach can assess the salient relationship of each region and find the salient objects which is the most interested one.

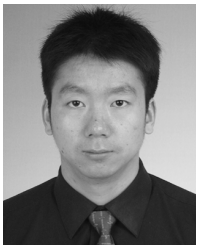
In addition, there are still some issues that need to be studied. When we change the salient feature maps to other validated salient features, a few of the measured images do not perform as well as the traditional methods. The neural network may selectively ignore the influence of the additional feature maps during the detecting. This induces that the proposed approach may be repressed in some situations. Besides, how to evaluate the effect of the more complex feature maps is a valuable problem which to be explored.

REFERENCES

[1] E. Ahn, J. Kim, L. Bi, A. Kumar, C. Li, M. Fulham, and D. D. Feng, "Saliency-based lesion segmentation via background detection in dermoscopic images," *IEEE J. Biomed. Health Inform.*, vol. 21, no. 6, pp. 1685–1693, Nov. 2017.

[2] S. Avidan and A. Shamir, "Seam carving for content-aware image resizing," *Trans. Graph.*, vol. 26, no. 3, p. 10, 2007.
[3] V. Navalpakkam and L. Itti, "An integrated model of top-down and bottom-up attention for optimizing detection speed," in *Proc. CVPR*, 2006, pp. 2049–2056.
[4] T. Durand, T. Mordan, N. Thome, and M. Cord, "WILDCAT: Weakly supervised learning of deep convnets for image classification, pointwise localization and segmentation," in *Proc. CVPR*, 2017, pp. 5957–5966.
[5] 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.
[6] W. Wang, J. Shen, and F. Porikli, "Saliency-aware geodesic video object segmentation," in *Proc. CVPR*, 2015, pp. 3395–3402.
[7] B. Mahasseni, M. Lam, and S. Todorovic, "Unsupervised video summarization with adversarial LSTM networks," in *Proc. CVPR*, 2017, pp. 2982–2991.
[8] T. Liu, H. Duan, Y. Shang, Z. Yuan, and N. Zheng, "Automatic salient object sequence rebuilding for video segment analysis," *Sci. China Inf. Sci.*, vol. 61, no. 1, 2018, Art. no. 012205.
[9] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 20, no. 11, pp. 1254–1259, Nov. 1998.
[10] S. Singh and R. Srivastava, "A novel probabilistic contrast-based complex salient object detection," *J. Math. Imag. Vis.*, vol. 61, no. 7, pp. 990–1006, 2019.
[11] X. Zhang, T. Wang, J. Qi, H. Lu, and G. Wang, "Progressive attention guided recurrent network for salient object detection," in *Proc. CVPR*, 2018, pp. 714–722.
[12] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. CVPR*, 2013, pp. 1155–1162.
[13] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. CVPR*, 2013, pp. 2083–2090.
[14] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. CVPR*, 2014, pp. 2814–2821.
[15] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proc. IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.
[16] G. Li, Y. Xie, L. Lin, and Y. Yu, "Instance-level salient object segmentation," in *Proc. CVPR*, 2018, pp. 247–256.
[17] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. CVPR*, Jun. 2016, pp. 678–686.
[18] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. ICCV*, 2017, pp. 202–211.
[19] J. Redmon and A. Farhadi, "YOLOv3: An incremental improvement," Apr. 2018, *arXiv:1804.02767*. [Online]. Available: https://arxiv.org/abs/1804.02767
[20] Y. Tian, G. Yang, Z. Wang, H. Wang, E. Li, and Z. Liang, "Apple detection during different growth stages in orchards using the improved YOLO-V3 model," *Comput. Electron. Agricult.*, vol. 157, pp. 417–426, Feb. 2019.
[21] G. Li and Y. Yu, "Visual saliency based on multiscale deep features," in *Proc. CVPR*, Jun. 2015, pp. 5455–5463.
[22] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. CVPR*, Jun. 2014, pp. 280–287.
[23] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. CVPR*, Jun. 2013, pp. 3166–3173.
[24] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. ECCV*, 2016, pp. 825–841.
[25] G. Li and Y. Yu, "Deep contrast learning for salient object detection," in *Proc. CVPR*, 2016, pp. 478–487.
[26] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. ICCV*, 2017, pp. 212–221.
[27] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," in *Proc. CVPR*, 2017, pp. 3203–3212.
[28] W. Wang, J. Shen, X. Dong, and A. Borji, "Salient object detection driven by fixation prediction," in *Proc. CVPR*, 2018, pp. 1711–1720.
[29] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proc. ECCV*, 2018, pp. 234–250.
[30] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, and A. Torralba, "Learning deep features for discriminative localization," in *Proc. CVPR*, 2015, pp. 2921–2929.
[31] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. CVPR*, 2017, pp. 618–626.
[32] J. Zhang, B. Li, Y. Dai, F. Porikli, and M. He, "Integrated deep and shallow networks for salient object detection," in *Proc. ICIP*, 2017, pp. 1537–1541.
[33] J.-X. Zhao, J. Liu, D.-P. Fan, Y. Cao, J. Yang, and M.-M. Cheng, "EGNet: Edge guidance network for salient object detection," in *Proc. ICCV*, 2019, pp. 1–10.
[34] J. Su, J. Li, Y. Zhang, C. Xia, and Y. Tian, "Selectivity or invariance: Boundary-aware salient object detection," in *Proc. CVPR*, 2019, pp. 1–10.

QIXIN CHEN received the bachelor's degree from the School of Information Engineering, Capital Normal University, where he is currently pursuing the master's degree with the School of Information Engineering. His research interests include computer vision, image processing based on deep learning, and visual saliency detection of images.



TIE LIU received the B.S., M.S., and Ph.D. degrees from Xi'an Jiaotong University, in 2001, 2004, and 2007, respectively. He is currently an Associate Professor with the College of Information Engineering, Capital Normal University. His research interests include machine learning, pattern recognition, multimedia computing, computer vision, data analysis, and mining.

YUANYUAN SHANG received the Ph.D. degree in electronic engineering from the Chinese Academy of Sciences, Beijing, China, in 2005. She is currently a Professor with the College of Information Engineering, Capital Normal University. Her research interests include image analysis and computer vision.

ZHUHONG SHAO received the Ph.D. degree in computer science and technology from Southeast University, Nanjing, China, in 2015. He is currently a Lecturer with the College of Information Engineering, Capital Normal University. His research interests include image analysis and pattern recognition.

HUI DING received the Ph.D. degree from the Beijing Institute of Technology, China, in 2006. She is currently an Associate Professor with the College of Information Engineering, Capital Normal University. Her research interests include image/video processing, especially video surveillance, and medical image analysis.