Edge Preserving and Multi-Scale Contextual Neural Network for Salient Object Detection

Xiang Wang, Huimin Ma, *Member IEEE*, Xiaozhi Chen, and Shaodi You

***Abstract*— In this paper, we propose a novel edge preserving and multi-scale contextual neural network for salient object detection. The proposed framework is aiming to address two limits of the existing CNN based methods. First, region-based CNN methods lack sufficient context to accurately locate salient object since they deal with each region independently. Second, pixel-based CNN methods suffer from blurry boundaries due to the presence of convolutional and pooling layers. Motivated by these, we first propose an end-to-end edge-preserved neural network based on Fast R-CNN framework (named *RegionNet*) to efficiently generate saliency map with sharp object boundaries. Later, to further improve it, multi-scale spatial context is attached to *RegionNet* to consider the relationship between regions and the global scenes. Furthermore, our method can be generally applied to RGB-D saliency detection by depth refinement. The proposed framework achieves both clear detection boundary and multi-scale contextual robustness simultaneously for the first time, and thus achieves an optimized performance. Experiments on six RGB and two RGB-D benchmark datasets demonstrate that the proposed method achieves state-of-the-art performance.**

***Index Terms*— Salient object detection, edge preserving, multi-scale context, RGB-D saliency detection, object mask.**

I. Introduction

SALIENT object detection, which aims to detect object that most attracts people's attention through out an image, has been widely exploited in recent years. It has also been widely utilized for many computer vision tasks, such as semantic segmentation [1], object tracking [2], [3] and image classification [4], [5].

Traditional saliency methods aim to generate a heat map which gives each pixel a relative value of its level of saliency [6]–[8]. In recent years, the fashion moves to salient object detection which generates pixel-wise binary label for salient and non-salient objects [9]–[11]. In comparing with the heat map, the binary label would further benefit segmentation based applications such as semantic segmentation [1], and thus attracts more attention.

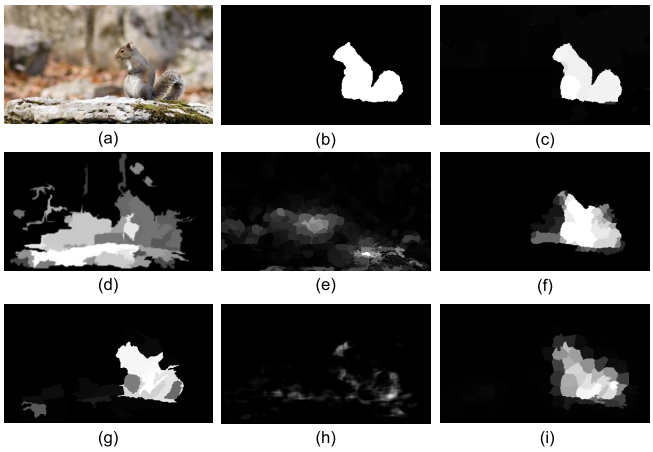


Fig. 1. Saliency map of an image with low-contrast. Previous methods fail to distinguish the object from the confusing background. Our method detect salient object with fine boundaries by taking advantages of regions and multi-scale context. (a) image, (b) groundtruth, (c) our proposed *RexNet*. (d, e) traditional methods: RC [10] and HDCT [17], (f, g) region-based CNN methods: LEGS [18] and MC [19], (h, i) pixel-based CNN methods: DISC [20] and DS [21].

To achieve a high accuracy for binary labeling, there are mainly two requirements: first, multi-scale contextual reliability; and second, sharp boundary between salient and non-salient objects. The contextual reliability aims to model the relationship between regions and global scenes to determine which object is salient. And the clear boundary aims to separate the salient object and background clearly and to highlight the whole object uniformly.

Unfortunately, none of the existing methods achieve both requirements simultaneously. Traditional bottom-up methods mainly rely on priors or assumptions and hand-crafted features. For example, center-surround difference [6], [12], uniqueness prior [13], [14] and backgroundness prior [15], [16]. These methods can not consider high-level semantic contextual relations and do not achieve a satisfying accuracy.

Recently, the deep Convolutional Neural Network (CNN) has attracted wide attention for its superior performance. CNN based methods can be divided into region-based networks and pixel-based networks. Region-based methods aim to extract features of each region (or patch), and then predict its saliency score. However, existing region-based methods lack of

Manuscript received February 27, 2017; revised August 7, 2017; accepted September 16, 2017. Date of publication September 26, 2017; date of current version October 17, 2017. This work was supported in part by the National Natural Science Foundation of China under Grant 61171113 and in part by the National Key Basic Research Program of China under Grant 2016YFB0100900. The associate editor coordinating the review of this manuscript and approving it for publication was Prof. Jianfei Cai. *(Corresponding author: Huimin Ma.)*

X. Wang, H. Ma, and X. Chen are with the Tsinghua National Laboratory for Information Science and Technology, Department of Electronic Engineering, Tsinghua University, Beijing 100084, China (e-mail: wangxiang14@mails.tsinghua.edu.cnl; mhmpub@tsinghua.edu.cn; chenxz12@mails.tsinghua.edu.cn).

S. You is with Data61, CSIRO, and Australian National University, Canberra ACT 0200, Australia (e-mail: Shaodi.You@data61.csiro.au).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2017.2756825

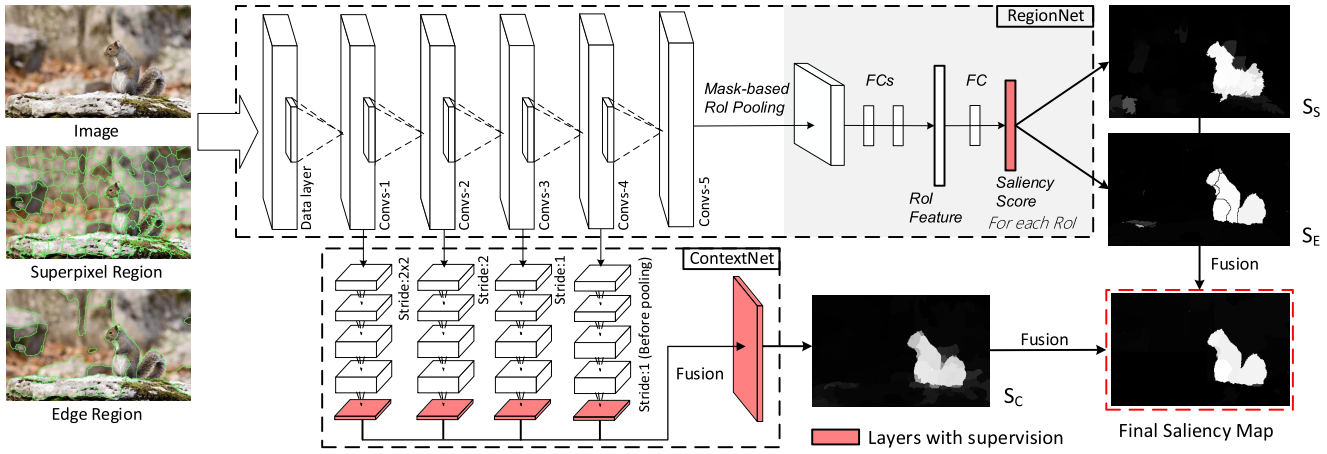


Fig. 2. Architecture of the proposed *RexNet*. The network is composed by two components: *RegionNet* and *ContextNet*. Image is first segmented into regions using superpixel and edges. *RegionNet* predicts saliency score of regions and forms saliency maps S_S and S_E. At the same time, *ContextNet* extracts multi-scale spatial context and fuse them to get saliency map S_C. These three saliency maps are fused to get the final saliency map.

representing context information to model the relationship between regions and global scenes. Because of this, it may have false detection results when the scene is complex or the object is composed by several different parts, which limits their performance (Fig. 1). On the other hand, existing pixel-based CNN methods lack the ability to produce clear boundary between salient and non-salient objects, due to the presence of convolutional and pooling layers, and they only achieve partial contextual reliability. This limits the performance of pixel-based methods (Fig. 1).

In this paper, we propose a novel edge preserving and multi-scale contextual network for salient object detection. The proposed framework achieves both clear boundary and multi-scale contextual robustness simultaneously for the first time. As illustrated in Fig. 2, the proposed structure, named *RexNet*, is mainly composed by two parts, the *REgionNet* and the *conteXtNet*. First, the *RegionNet* is inspired by the Fast R-CNN framework [22]. Fast R-CNN is recently proposed for object detection and achieves superior performance because the convolutional features of entire image are shared and features of each patch (or RoI) are extracted via the RoI pooling layer. We extend Fast R-CNN to salient object detection by introducing mask-based RoI pooling and formulating salient object detection as a binary region classification task. The image is first segmented into regions and are used as input of *RegionNet*, the *RegionNet* then predicts saliency score of each region end-to-end to form saliency map of the entire image. Since the regions are segmented by edge-preserved methods, saliency map generated by our network is naturally with sharp boundaries.

Second, the *ContextNet* aims to provide strongly reliable multi-scale contextual information. Different from most previous works which consider context by expanding region window at a certain layer, in this paper, we consider to model context via multiple spatial scales. This is based on the observation that different layers of CNN represent different levels of semantic [23], [24], considering context of different levels may be more sufficient. We achieve this by taking advantages of dense image prediction. For all max-pooling layers of *RegionNet*, we attach multiple convolutional layers to predict saliency map of different levels. Then all levels of saliency map are fused with *RegionNet* to generate the final saliency map. Our method generates saliency map with accurate location while keeping fine object boundaries.

Other than the effectiveness, our proposed frameworks is efficient, since we take advantages of regions by extending the efficient Fast R-CNN framework, which predicts saliency score of regions by only one forwarding. We also extend our method to RGB-D saliency by applying depth refinement. Experiments on 2 RGB-D benchmark datasets demonstrate that the proposed *RexNet* outperforms other methods by a large margin.

The main contributions of this paper are three-fold. First, we proposed *RegionNet* which generates saliency score of regions efficiently and preserves object boundaries. Second, multi-scale spatial context is considered and attached to *RegionNet* to boost salient object detection performance. Third, we extend our method to RGB-D saliency datasets and use depth information to further refine saliency maps.

The rest of this paper is organized as follows. Section II discusses related work. Section III and Section IV introduce the details of the proposed *RegionNet* and *ContextNet* correspondingly. Section V describes the training details of the proposed network. Section VI introduces our extension to RGB-D salient object detection. Section VII shows the experimental results and comparison with state-of-the-art methods. And conclusion is made in Section VIII.

II. Related Work

In this section, we introduce traditional salient detection methods and the recent CNN based methods. In addition, we also introduce some related works that integrate multi-scale context information and some topics related to salient object detection.

A. Traditional Methods

Salient object detection was first exploited by Itti *et al.* [6], and later attracted wide attention in the computer vision society. Traditional methods mostly rely on prior assumptions and most are un-supervised. Center-surround difference which assumes that salient regions differs from their surrounding regions is an important prior in early research. Itti *et al.* [6] first proposed center-surround difference at different scales to compute saliency. Liu *et al.* [12] propose center-surround histogram which defines saliency as the difference between center region and its surrounding region. Li *et al.* [25] propose cost-sensitive SVM to learn and discover salient regions that are different from their surrounding regions. These methods cannot provide sharp boundary for salient region because they are based on rectangle regions, which is only able to generate coarse and blurry boundary.

While center-surround difference considers local contrast, it does not take into consideration of global contrast. Global contrast based methods are later proposed, *e.g.*, Cheng *et al.* [10] and Yan *et al.* [26]. In [10], image is first segmented into superpixels. Then saliency value of each region is defined as the contrast with all other regions. The contrast is weighted by spatial distance so that nearby regions have greater impact on it. To deal with objects with complex structures, Yan *et al.* [26] propose a hierarchical model which analyzes saliency cues from multiple scales based on local contrast and then infers the final saliency values of regions by optimizing them in a tree model. Following them, many methods utilizing bottom-up priors are proposed, readers are encouraged to find more details in a recent survey paper by Borji *et al.* [11].

B. CNN Based Methods

Deep Convolutional Neural Network (CNN) has attracted a lot of attention for its outstanding performance in representing high-level semantic. Here, we mention are few representative work. These work can be divided into two categories according to their treatment of input images: region-based methods and pixel-based methods. Region-based methods formulate salient object detection as a region classification task, namely, extracting features of regions and predict their saliency score. While pixel-based methods directly predict saliency map pixels-to-pixels with CNN.

1) Region-Based Methods: Wang *et al.* [18] propose to detect salient object by integrating both local estimation and global search with two trained networks DNN-L and DNN-G. Zhao *et al.* [19] consider global and local context by putting a global and a closer-focused superpixel-centered window to extract features of each superpixel, respectively, and then combine them to predict saliency score. Li *et al.* [27] propose multi-scale deep features by extracting features of each region at three scales and then fuse them to generate its saliency score. These works are region-based which focused on extracting features of regions and fuse larger scale of regions as context to predict saliency score of each region. These fusions are mostly applied at only one layer and does not achieve a optimal performance. In addition, the networks extract features of one region for each forwarding which is very time-consuming.

2) Pixel-Based Methods: Recently, CNN has also been applied to pixels-to-pixels dense image prediction, such as semantic segmentation and saliency prediction. Long *et al.* [28] propose fully convolutional networks which is trained end-to-end and pixels-to-pixels by introducing fully convolutional layers and a skip architecture. Chen *et al.* [20] propose a coarse-to-fine manner in which the first CNN generates coarse map using the entire image as input and then the second CNN takes the coarse map and local patch as input to generate fine-grained saliency map. Li *et al.* [21] propose a multi-task model based on fully convolutional network. In [21], saliency detection task is in conjunction with object segmentation task, which is helpful for perceiving objects. A Laplacian regularized regression is then applied to refine saliency map. However, while end-to-end dense saliency prediction is efficient, the resulting saliency maps are coarse and with blurry object boundaries due to the presence of convolutional layers with large receptive fields and pooling layers.

C. RGB-D Salient Object Detection

RGB-D saliency is an emerging topic and most RGB-D saliency methods are based on fusing depth priors with RGB saliency priors. Ju *et al.* [29] propose RGB-D saliency method based on anisotropic center-surround difference, in which saliency is measured as how much it outstands from surroundings. Peng *et al.* [30] propose depth saliency with multi-contextual contrast and then fuse it with appearance cues via a multi-stage model. Ren *et al.* [31] propose normalized depth prior and global-context surface orientation prior based on depth information and then fuse them with RGB region contrast priors. Depth contrast may cause false positives in background region, to address it, in [32], Feng *et al.* propose local background enclosure feature based on the observation that salient objects tend to be locally in front of surrounding regions. To the best of our knowledge, existing RGB-D salient object detection are all using hand-crafted features and the performance is not optimized.

D. Multi-Scale Context

Multi-scale context has been proved to be useful for image segmentation task [19], [27], [33], [34]. Hariharan *et al.* [33] proposed hypercolumns for object segmentation and fine-grained localization, in which they defined "hypercolumn" at a given input location as the outputs of all layers at that location. Features of different layers are combined and then be used for classification. Zhao *et al.* [19] proposed multi-context network which extracts features of a given superpixel at global and local scale, and then predict saliency value of that superpixel. Li *et al.* [27] proposed to extract features at three scales: bounding box, neighbourhood rectangular and the entire image. Liu *et al.* [34] proposed to use recurrent convolutional layers (RCLs) [35] iteratively to integrate context information and to refine saliency maps. At each step, the RCL

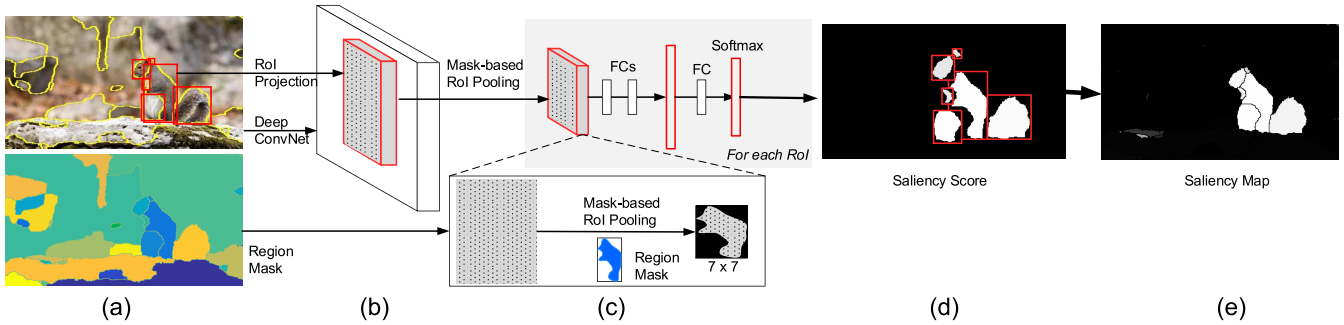


Fig. 3. Pipeline of *RegionNet*. We extend the Fast R-CNN framework for saliency detection. (a) Image is first segmented into regions and the region mask which records the index of regions is also generated. For each region, we use its external rectangle as RoI. Note that, for clarity, we only show RoIs of salient objects, the background regions are omitted. (b) All RoIs are put into the convolutional networks, and (c) at the RoI pooling layer, the mask-based RoI pooling is applied to extract features inside region mask. In this way, the features of irregular region can be extracted. (d) With this mask-based pooling, the framework predicts saliency score of regions end-to-end, and (e) to form the saliency map of the entire image.

takes coarse saliency map from last step and feature map at lower layer as input to predict a finer saliency map. In this way, context information is integrated iteratively and the final saliency map is more accurate than that predicted from global context.

The proposed *ContextNet* differs from those at two aspects. First, the *ContextNet* is a holistically-nested architecture [36] which predicts saliency map at each branch and fuse them finally. Second, we propose *EdgeLoss* as a supervision which makes the boundary of segmentation result more clear.

E. Fixation Prediction and Semantic Segmentation

Fixation prediction [6]–[8], [37] aims to predict the regions people may pay attention to, and semantic segmentation [28], [38] aims to segment objects of certain classes in images. They are topics related to salient object detection, but they also have significant differences. Fixation prediction aims to predict *regions* which most attract people's attention, while salient object detection focuses on segmenting the most attractive *objects*. For semantic segmentation, saliency detection is a class-agnostic task, whether an object is salient or not is largely depend on its surroundings, while semantic segmentation mainly focuses on segmentation objects of certain classes (*e.g.* 20 classes in PASCAL VOC dataset). So compared with semantic segmentation, context information is more important for saliency detection, and this is the main motivation of our *ContextNet*.

III. RegionNet: Edge Preserving Neural Network for Salient Object Detection

A. Motivation

In this paper, we aim to propose a unified framework which can preserve object boundaries and take multi-scale spatial context into consideration. To preserve object boundaries, we propose an effective network, named *RegionNet*, which generates saliency score of each region end-to-end (Fig. 3). Different from previous region-based methods [18], [19], [27], we extend the efficient Fast R-CNN framework [22] for salient object detection for the first time. On the other hand, previous works consider context mainly by expanding window of region or using entire images at a certain data or feature layer. In this paper, we consider context at multiple layers and using dense saliency prediction framework to generate saliency maps to complement *RegionNet*. The architecture of the proposed framework is shown in Fig. 2.

In this section, we first introduce the idea of edge-preserving saliency detection based on a CNN network. This idea is previously appeared in our conference paper [39]. In section IV, we extend this idea with consideration of multi-scale spatial context.

B. RegionNet

In this section, we introduce *RegionNet* which takes advantage of CNN for high effectiveness and high efficiency. More importantly, it takes advantage of region segmentation which enables clear detection boundary and further improves the accuracy.

1) Network Architecture: We extend original Fast R-CNN [22] structure for end-to-end saliency detection. Fast R-CNN is an efficient and general framework in which the convolutional layers are shared on the entire image and the feature of each region is extracted by the RoI pooling layer. However, to the best of our knowledge, Fast R-CNN is only used for object detection and classification but not for saliency. Namely, the result of Fast R-CNN is bounding box but not pixel-wise. In this paper, we make the modification to enable edge preserving saliency by introducing mask-based RoI pooling. Different from previous region-based methods which deal with each region of an image independently, our proposed Fast R-CNN structure processes all regions end-to-end and with the entire image considered.

2) Detection Pipeline: As illustrated in Fig. 3, first, given an image, we segment it into regions using superpixel and edges. And for each region, we use its external rectangle as proposal (or RoI) and use it as input of Fast R-CNN framework similar with object detection tasks. We also generate a region mask with the same size of image to record the region index for each pixel and then downsample it by 16 times and put it into the RoI pooling layer.

Then, at the RoI pooling stage, features inside each RoI ($h \times w$) are pooled into a fixed scale $H \times W$ (7×7 in our work).

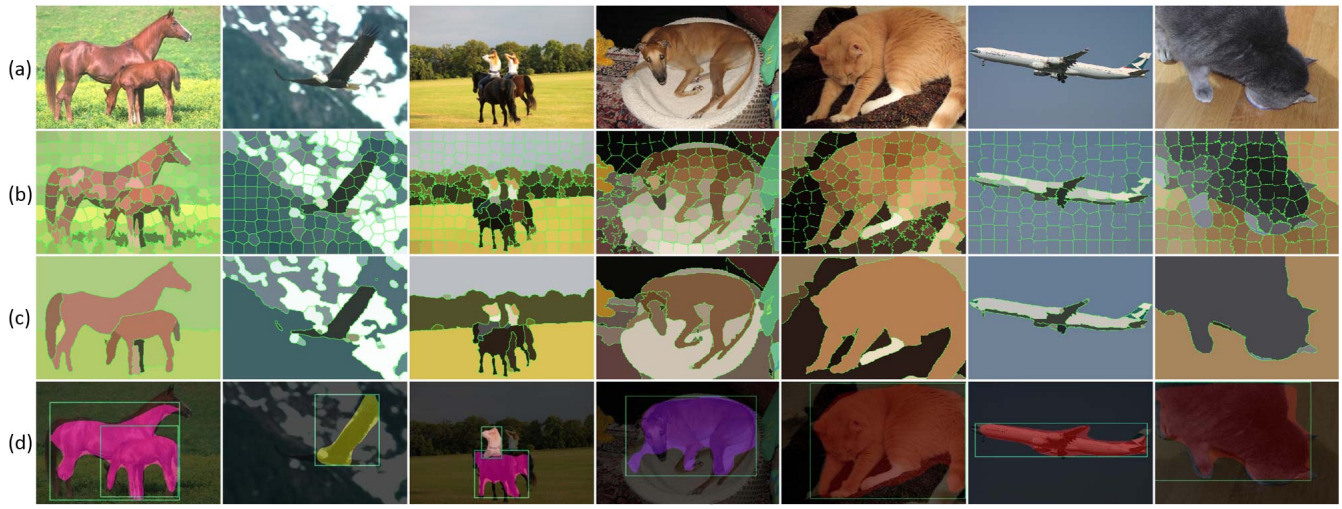


Fig. 4. (a) images, (b) and (c) superpixel regions and edge regions. Pixels in each region are replaced with their mean color, (d) masks generated by MNC [40]. (i) We can see that edges divide images into fewer regions than superpixels and thus preserving more compactness of objects, which is helpful for saliency prediction. (ii) The superpixels and edges regions achieve higher boundary accuracy than masks generated by MNC [40]. Best viewed in color.

So each sub-window with scale $h/H \times w/W$ is converted to one value with max-pooling. To extract feature of irregular pixel-wise RoI region, we only pool features inside its region mask while leaving others as 0. The process of the proposed mask-based RoI pooling is formulated as following. For region with index i, and a certain sub-window as SW_j, we denote region mask as M, features before pooling as F, the pooled feature at sub-window SW_j as P_j, then

$$P_j = \begin{cases} \max\limits_{\{k|k\in SW_j, M_k=i\}} F_k & i \in M(SW_j), \\ 0 & i \notin M(SW_j). \end{cases} \tag{1}$$

With this mask-based pooling, features of each region are extracted and the edge information is also preserved.

Last, by considering salient object detection as a binary classification problem, the network generates saliency score of regions to form the saliency map of entire image end-to-end.

Note that, in our work, to segment image into regions, besides superpixel, we also consider larger scale regions which are segmented by edges (denoted as edge regions). This is based on the observation that when an object is segmented into dozens of superpixels, it will be difficult to uniformly highlight the whole object. The edge regions can preserve more compactness of objects and thus may be more effective. Recent advances in edge detection have achieved highly satisfactory performance which makes it practical to use edge information to help better detect salient objects. In our work, we use HED method of Xie *et al.* [36] to get object edges and then thinning them using method of Dollar *et al.* [41]. The superpixel is segmented using SLIC algorithm [42].

Some examples of superpixel regions and edge regions are shown in Fig. 4. We can see that edges segment image into fewer regions and better preserves compactness of object. For region-based methods, this will help improve the final performance and since the number of regions is smaller, it also reduces computation cost. Considering the fault-tolerant capability, namely, misclassification of edge regions may decrease performance largely, the superpixel regions are also used in our method. These two scales regions are complementary since superpixel regions can generate results with high resolution and edge regions can preserve more compactness of objects.

Note that the similar idea of mask-based RoI pooling has also been applied in MNC [40] for semantic segmentation. However, we have much difference. In [40], the masks were generated by the multi-task network and they are continuous values in [0, 1]. The masked feature is the element-wise product of features and masks. While in our work, the masks are got by segmenting images into regions with superpixels [42] and edges [36], they are binary and the mask-based RoI pooling is to extract features inside the masks. The SLIC algorithm [42] for generating superpixels has strong ability to adhere to image boundaries, so its boundary accuracy is quite good. The HED [36] network is designed for edge detection, the boundary accuracy is much better than multi-task networks in [40]. So the masks of our method has higher boundary accuracy compared with MNC [40]. Some examples are shown in Fig. 4.

We denote the saliency map generated by *RegionNet* with superpixel regions and edge regions as S_S and S_E, respectively. We have shown in our previous conference paper [39] that S_E outperforms most previous works, and the combination of S_E and S_S achieves better performance, which shows the effectiveness of edge regions and the combination with superpixel regions. More detailed experimental results are shown in Section VII.

IV. ContextNet: Multi-Scale Contextual Neural Network for Salient Object Detection

In this section, we introduce the extension of the proposed method by utilizing multi-scale context. In Section IV-A, we first introduce the motivation for multi-scale context, after that, in Section IV-B, we introduce the architecture of the

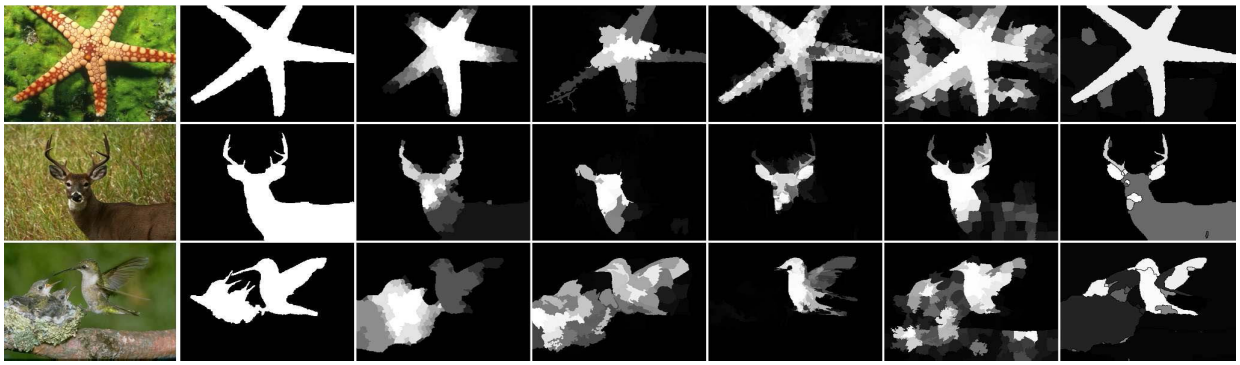


Fig. 5. Results of previous region-based methods and our S_S and S_E. We can see that misclassification of regions has a great impact on the final performance and most regions are assigned to near either 0 or 1, with few intermediate values. These will limit the precision at high recall when thresholding.

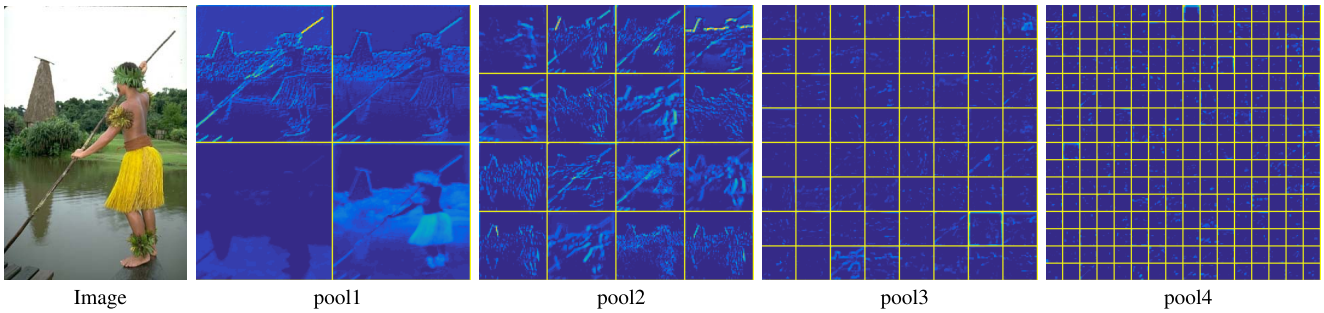


Fig. 6. Visualization of features in different layers of *RegionNet*. For a test image, we forward it in our trained *RegionNet*, and then we extract features of the first four pooling layers and show each channel of them. Different layer represents different level of semantic. Best viewed in color.

proposed multi-scale contextual network. In Section IV-C, we introduce the loss function for supervising the *ContextNet*, and in Section IV-D, we introduce deep supervision to accelerate convergence and improve prediction performance.

A. Motivation

Salient object detection is a class-agnostic task, whether a region is salient or not is largely depend on its surroundings, *i.e.*, context. While the *RegionNet* we proposed can generate saliency map with well preserved boundary, it lacks of context information. In addition, region-based CNN methods [18], [19], [27] suffer from some common drawbacks. First, region-based methods are based on binary region classification, misclassification of regions will cause large false detection. Second, solving binary classification problem with huge amount of images using CNN causes the classification results to be extremely separated to either 0 or 1, thus saliency map is not smooth. These two issues will limit the precision at high recall. Fig. 5 shows some results of previous region-based CNN methods and our S_S and S_E.

As explored in previous works [23], [24], features in different layers of CNN has different properties and represent different levels of semantic. So fusing context from multiple layers may be more sufficient. Fig. 6 shows the visualization example of features in the first four pooling layers of *RegionNet*. We can see that shallow layers mainly focus on bottom features, such as contour, and deep layers focus on more abstract high-level features. Based on these observations, in this paper, we consider context information by introducing multi-scale contextual layers, named *ContextNet*, to address the issues mentioned above and to complement *RegionNet*.

B. Network Architecture

The architecture of our proposed network is shown in Fig. 2. Based on the *RegionNet*, we propose to use multi-scale dense image prediction method to model the relationship between regions and the global scenes at multiple levels. For all max pooling layers (except the RoI pooling layer) of *RegionNet*, we attach five convolutional layers (called as branch) to predict saliency maps of different levels. The first three layers of each branch are with 3×3 convolutional filters and 64, 64, 128 channels, and the dilated convolution [38] is also applied to increase the receptive field. The last two layers are fully convolutional layers with 128 and 1 channels.

Experimental results in [28] have demonstrated that denser prediction map has better performance. Following that, we propose to generate saliency map with one eighth scale of the original input images. So we set the stride of each branch as 4, 2, 1, 1, respectively. Note that the last branch is connected to the convolution layer before the fourth max-pooling layer, *i.e.,* conv4_3 in VGG16 [43], so output of all branches have the same dimensions. The outputs of all branches are then fed into fully convolutional layers which learn the combination weights to generate saliency map S_C. The final saliency map S is then got by fusing S_S, S_E, and S_C via a fully

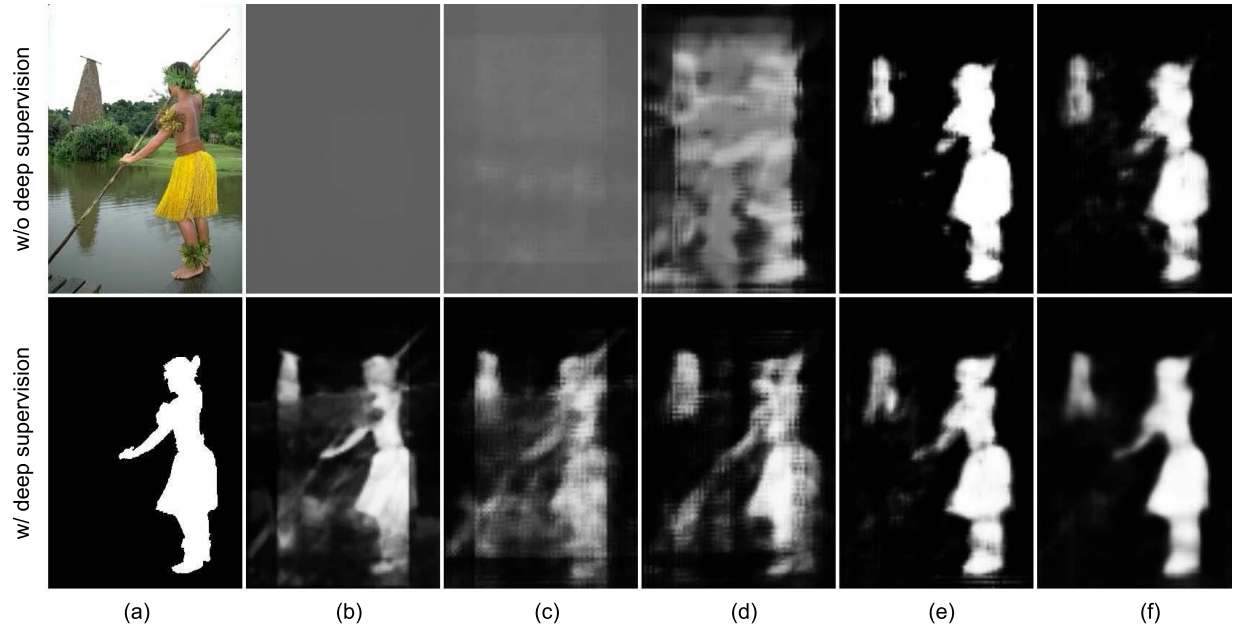


Fig. 7. Effect of deep supervision. From left to right are image and ground truth, results of 4 branches, and fusion of all branches. The first row shows results without deep supervision and the second row shows results with deep supervision. Without deep supervision, the first and second branch learn almost nothing in our network due to the heavy bias.

convolutional layer.

$$S = Fusion(S_S, S_E, S_C). \tag{2}$$

C. Loss

We assume that the training data, $\mathcal{D} = \{(X_i, T_i)\}_{i=1}^{N}$, consists of N training images and groundtruth. Our goal is to train a convolutional network $f(X;\theta)$ to predict saliency map of a given image. We define two kinds of loss for *ContextNet* to generate saliency map with high accuracy and clear object boundary.

The first *Loss* is common used Cross Entropy Loss $\mathcal{L}_\mathcal{C}$, which aims to make the output saliency map $f(X;\theta)$ consistent with the groundtruth T.

$$\mathcal{L}_\mathcal{C} = -\frac{1}{N}\sum_{i=1}^{N}[T_i log(f(X_i;\theta)) + (1-T_i) log(1-f(X_i;\theta))]. \tag{3}$$

The second *Loss* is Edge Loss $\mathcal{L}_\mathcal{E}$ which aims to preserve edge and make the saliency map more uniform. Since we have segmented image into regions with edge-preserved methods, our assumption is that saliency map in the same region should share similar value, so that the final saliency map can also preserve edge and be more uniform. We average saliency map $f(X;\theta)$ in each region and marked the averaged map as $\bar{f}(X;\theta)$. The Edge Loss is defined as the $L2$ norm between saliency map $f(X;\theta)$ and the averaged map $\bar{f}(X;\theta)$.

$$\mathcal{L}_\mathcal{E} = \frac{1}{2N}\sum_{i=1}^{N} \| f(X_i;\theta) - \bar{f}(X_i;\theta) \|_2^2. \tag{4}$$

D. Deep Supervision

The proposed *ContextNet* comprises of a fusion layer which fuses the outputs of four branches. Supervision only in the last fusion layer may cause heavy bias, namely, some layers may not be optimized adequately. To address this issue, in this paper, we utilize deep supervision [36], [44] method, namely, outputs of all branches and their fusion result are also supervised. Fig. 7 shows the comparison of results with and without deep supervision. Without deep supervision, the network will be heavily biased towards some maps, and in extreme cases, some branches will learn nothing, *e.g.*, Fig. 7 (b) and (c). While with deep supervision, each branch learns and predicts saliency map with features at different scale, which accelerates convergence of the network and makes the final saliency map more precise.

V. Network Training

We implement our method using Caffe framework [45]. The training process consists of two stages. At the first stage, we fine-tune the *RegionNet* using weights pre-trained on ImageNet [46]. At the second stage, we fix the weights of *RegionNet* and then optimize the weights of the *ContextNet* using SGD procedure.

For the training of *RegionNet*, a region is considered as salient/background if more than 80% of its pixels are located inside/outside ground truth. The *RegionNet* formulates salient object detection as a binary classification problem and the loss function we used is softmax loss. Following previous works, we fine-tune our *RegionNet* based on VGG16 [43] which is pre-trained on ImageNet [46].

For the training of *ContextNet*, deep supervision is applied to accelerate convergence and to improve the final performance.

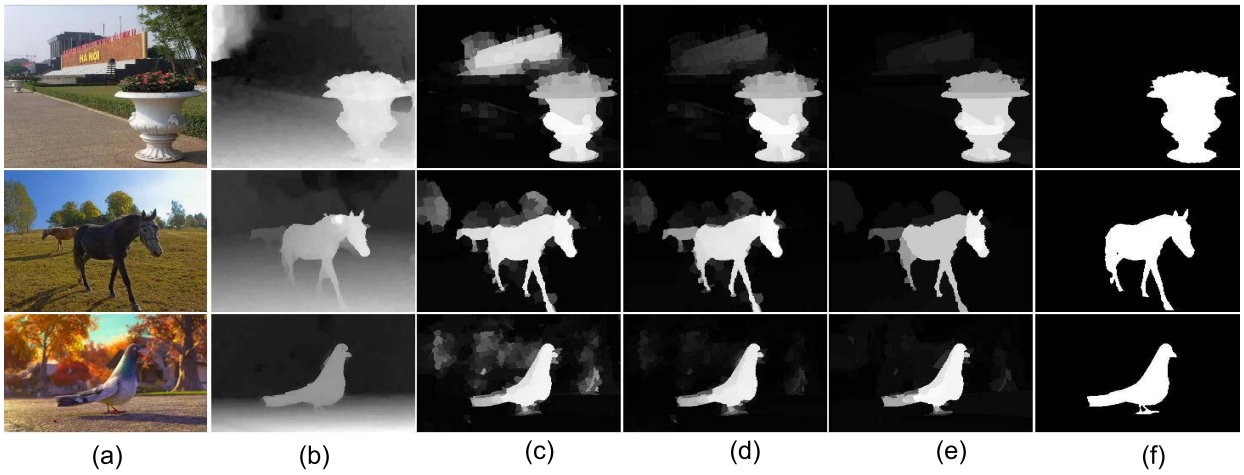


Fig. 8. The process of depth refinement. (a) image, (b) depth, (c) saliency map of our method using RGB data (S_0), (d) with the position prior, the background noise is strongly suppressed (S_1), and (e) with the local compactness prior, the background is further suppressed and the result map is more uniform (S_2), (f) groundtruth.

VI. Extension to RGB-D Salient Object Detection

Depth information is an important cue for salient object detection, especially for images with complex scenes. In this paper, we apply depth information to further improve the performance by extending our framework to RGB-D saliency datasets.

For RGB-D datasets, a simple idea is to train our network using RGB-D data directly. However, it suffers from two problems. First, our network is pre-trained on ImageNet [46], it is unreasonable to fine-tune it using RGB-D data. Second, the image number of existing RGB-D saliency dataset is too small to well train a network. So in this paper, we propose to first generate saliency map using RGB data, and then refine it with depth information.

We propose two efficiency priors based on our observations: position prior and local compactness prior. For position prior, in most scenes, the salient object is located at the most front position. For local compactness prior, regions with similar depth, appearance and position should share similar saliency value.

We denote saliency map generated by our network as S_0. For position prior, we directly multiply S_0 by depth D using a sigmoid function and denote it as S_1,

$$S_1 = S_0 \times \frac{1}{1 + exp(-\sigma \times D)}, \tag{5}$$

in which the parameter σ is set to 5 empirically in our work. Note that we have transformed the depth similar with [29], in which the depth is rescaled to [0, 1] and pixels with shorter distance are attached with larger intensity.

For local compactness prior, saliency value of each region $S_2(i)$ is refined with their neighbor regions $\mathcal{N}(i)$ weighted by depth and appearance similarity.

$$S_2(i) = \sum_{j \in \mathcal{N}(i)} W(i, j) S_1(j), \tag{6}$$

with

$$W(i, j) = exp(-\frac{D(i, j)^2}{2\sigma_{dep}^2}) exp(-\frac{Col(i, j)^2}{2\sigma_{col}^2}), \tag{7}$$

in which $Col(i, j)$ denotes the Euclidean distance of RGB color. We set $\sigma_{dep} = 0.02$ and $\sigma_{col} = 5$ empirically in our work. Fig. 8 shows some examples of the depth refinement.

VII. Experiments

To evaluate the effectiveness of each component and study the performance of the proposed method, we conduct experiments on six RGB and two RGB-D benchmark datasets and compare our method with state-of-the-art methods quantitatively and qualitatively.

A. Setup

We randomly sample 4000 images from DUT-OMRON [47] dataset and 5000 images from MSRA10K [10], [12], [48] dataset as training set and then evaluate our method on the following six benchmark datasets: ECSSD [26], DUT-OMRON [47], JuddDB [49], SED2 [50], THUR15K [51] and Pascal-S [52]. Note that the DUT-OMRON has 5168 images and we only evaluate on the remaining 1168 images that are not included in the training set. We also evaluate our method on two benchmark RGB-D saliency datasets: RGBD1000 [30] and NJU2000 [29]. All results are got from the benchmark of Borji *et al.* [53] or generated using authors' code.

We evaluate the performance using precision-recall (PR) curves, F-measure and mean absolute error (MAE). The saliency maps are first normalized to [0, 255], and then the precision and recall are computed by binarizing them with 256 thresholds and comparing them with ground truth. The PR curves are computed by averaging them on each dataset. The F-measure considers both precision and recall which is computed as:

$$F_\beta = \frac{(1 + \beta^2) Precision \times Recall}{\beta^2 Precision + Recall}, \tag{8}$$

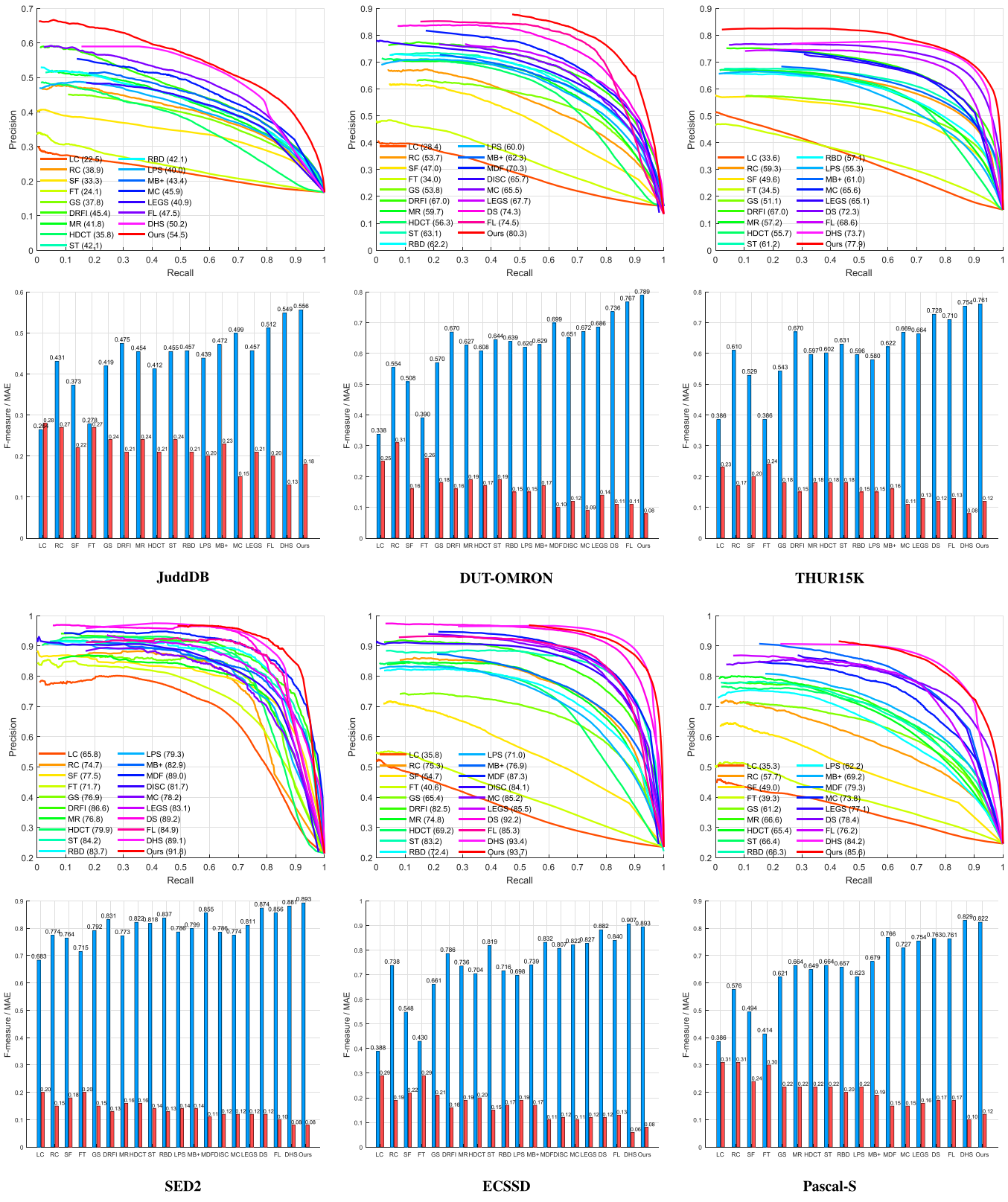


Fig. 9. Comparison with state-of-the-art methods on six benchmark datasets. For each dataset, the first row shows the PR curves and the second row shows the F-measure and MAE. The numbers in the PR curves denote the AUC. Best viewed in color.

we set $\beta^2 = 0.3$ as most previous works [10], [48] to emphasize the precision. The final F-measure is the maximal F_β computed by 256 precision-recall pairs in the PR curves [53]. The MAE directly measures the mean absolute difference between saliency map and ground truth,

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

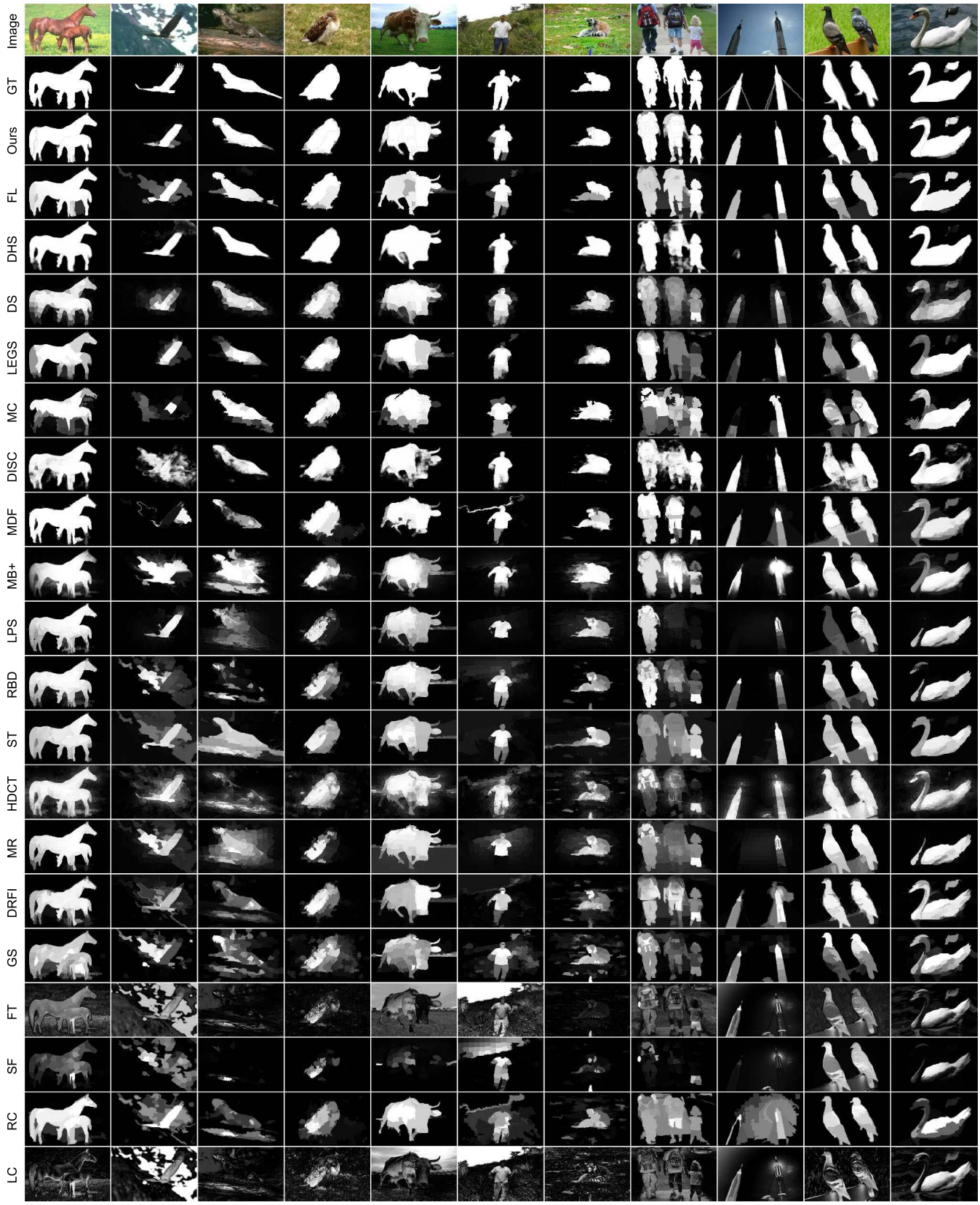


Fig. 10. Qualitative comparison with state-of-the-art methods. We can see that our method locates salient objects more accurately and preserves object boundaries better. Background noise is strongly suppressed and the objects are highlighted uniformly.

B. Comparison With State-of-The-Art Methods

We compare our method with state-of-the-art methods, including traditional methods: LC [9], RC [10], SF [54], FT [48], GS [15], DRFI [55] MR [47], HDCT [17], ST [56], RBD [16], LPS [57], MB+ [58], and CNN based methods: MDF [27], DISC [20], MC [19], LEGS [18], DS [21],

TABLE I
TRAINING DATA OF STATE-OF-THE-ART METHODS

Method	Training Data
MDF [27]	2,500 images from MSRA-5000
DISC [20]	9,000 images from MSRA10K
MC [19]	8,000 images from MSRA10K
LEGS [18]	3,000 images from the MSRA-5000 dataset and 340 images from the Pascal-S dataset. Both horizontal reflection and rescaling (5%) are applied
DS [21]	leave-one-out strategy, using other 7 datasets for training
DHSNet [34]	6,000 from MSRA10K and 3,500 from DUT-OMRON
OURS	4,000 from DUT-OMRON and 5,000 from MSRA10K

DHSNet [34] and our preliminary conference method FL [39]. For CNN-based methods, we also list the training data they used in Table I. MDF [27] uses less training data, DS [21] uses much more training data, and for other methods, we use comparable training data. Fig. 9 shows PR-curves, F-measure and MAE on six benchmark datasets. We can see that our method outperforms other methods and our preliminary conference method by a large margin. For the state-of-the art multi-scale method DHSNet [34], we achieve comparable performance. For PR curves, our method outperforms DHSNet on all datasets by 2.6% on average. For F-measure, our method outperforms DHSNet on JuddDB, THUR15K and SED2 datasets, but fails on ECSSD and Pascal-S dataset. For MAE, we are inferior to DHSNet by 0.026 on average.

Note that DS [21] is a multi-task framework which detects salient object and object boundaries simultaneously, our method outperforms DS [21] at all 6 datasets, especially on datasets with complex scenes, such as DUT-OMRON, JuddDB and Pascal-S, which shows that our method better takes advantages of edges. Note that our network is trained on parts of DUT-OMRON and MSRA10K dataset, we apply the trained network to other 5 datasets without fine-tuning, the results still outperform others by a large margin, which shows that our method has strong generalization ability. Fig. 10 shows the qualitative comparison with state-of-the-art methods, we can see that our method preserves edges well and suppresses most background noise.

C. Evaluation on RGB-D Saliency Datasets

We compare our method with state-of-the-art RGB-D saliency methods: ACSD [29], GP [31], LMH [30] and LBE [32]. Fig. 11 shows the comparison of PR-curves. Our method significantly outperforms other methods, especially in the region of high recall. The main reason of our performance is that our method can not only locate salient object accurately, but also preserve edges, thus saliency map of our method are with high precision and high recall. Fig. 12 also shows the qualitative comparison with state-of-the-art RGB-D methods.

D. Ablation Studies

In this subsection, we conduct experiments to verify the effectiveness of each component of our method.

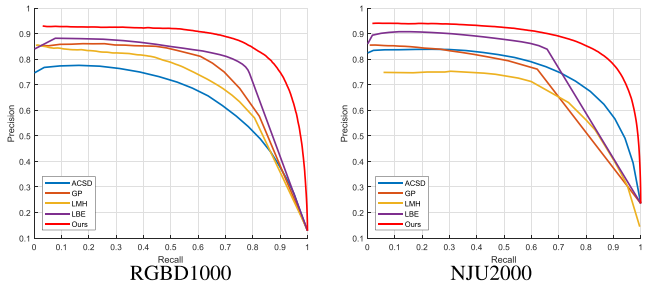


Fig. 11. Comparison with state-of-the-art methods on two benchmark RGB-D saliency datasets. Best viewed in color.

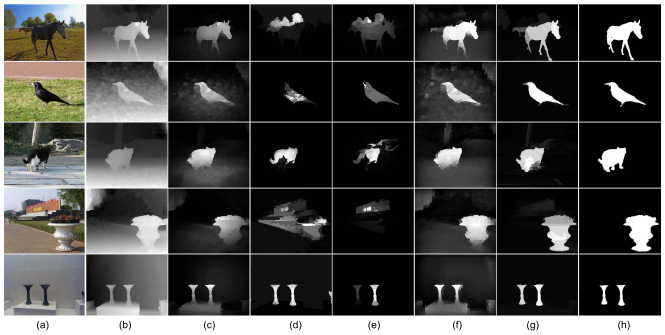


Fig. 12. Qualitative comparison with state-of-the-art methods on RGB-D datasets. Our method can not only locate salient object accurately, but also preserve edges, thus highlighting the whole object uniformly and suppressing background noise.

1) Network Components: We first evaluate the components of the proposed network by outputting the intermediate results of our network and analyzing their performance. Table II shows the comparison of all components: S_S, S_E, S_C and the final saliency map S on six benchmark datasets. To better demonstrate the comparison with numerical results, we use Area Under Curve (AUC) which measures the area under the PR-curve to represent PR-curve criterion. We can see that the final result S outperforms all components, which shows that all the components are complementary and our method is effective.

2) Branches of ContextNet: We evaluate the effectiveness of branches of *ContextNet*. Table III shows the results of each branch and the fusion results on six benchmark datasets. We can see that, commonly, the branches of deeper layers achieve better performance, and the final fusion result is the best, which demonstrates that our method makes full use of features at each branch.

3) Edge Loss: We evaluate the effectiveness of Edge Loss by comparing with networks without Edge Loss. Table IV shows the results of *ContextNet* on six benchmark datasets. With the Edge Loss, the performance is better since the Edge Loss can preserve edges better and so the saliency map of *ContextNet* are more uniform.

4) Comparison With Fusing Features: The proposed *ContextNet* fuses saliency maps of each branch to get the final result. To evaluate the effectiveness, we compare with method which fuses features to predict saliency map. We concatenate features of each branch to predict saliency map. Table V

TABLE II

EVALUATION OF ALL COMPONENTS ON SIX BENCHMARK DATASETS WITH F-MEASURE AND AUC. THE FINAL RESULT S ALWAYS PERFORMS BETTER THAN ALL COMPONENTS, WHICH SHOWS THAT ALL THE COMPONENTS ARE COMPLEMENTARY AND OUR METHOD IS EFFECTIVE

	JuddDB		DUT-OMRON		THUR15K		SED2		ECSSD		Pascal-S	
	F_β	AUC	F_β	AUC	F_β	AUC	F_β	AUC	F_β	AUC	F_β	AUC
S_S	0.490	0.457	0.722	0.720	0.706	0.710	0.849	0.851	0.851	0.901	0.768	0.806
S_E	0.515	0.464	0.771	0.728	0.734	0.696	0.882	0.861	0.858	0.864	0.789	0.802
S_C	0.534	0.508	0.762	0.770	0.721	0.717	0.877	0.883	0.874	0.914	0.799	0.836
S	**0.556**	**0.545**	**0.789**	**0.803**	**0.761**	**0.779**	**0.893**	**0.918**	**0.893**	**0.937**	**0.822**	**0.856**

TABLE III

RESULT OF EACH BRANCH AND THEIR FUSION IN *ContextNet*

	JuddDB		DUT-OMRON		THUR15K		SED2		ECSSD		Pascal-S	
	F_β	AUC	F_β	AUC	F_β	AUC	F_β	AUC	F_β	AUC	F_β	AUC
Branch 1	0.402	0.366	0.529	0.510	0.533	0.510	0.749	0.780	0.639	0.643	0.599	0.596
Branch 2	0.416	0.381	0.525	0.507	0.557	0.540	0.691	0.728	0.692	0.719	0.622	0.622
Branch 3	0.447	0.423	0.564	0.563	0.600	0.601	0.705	0.737	0.751	0.801	0.678	0.713
Branch 4	0.490	0.457	0.692	0.710	0.686	0.695	0.802	0.854	0.836	0.891	0.756	0.798
Fusion	**0.534**	**0.508**	**0.762**	**0.770**	**0.721**	**0.717**	**0.877**	**0.883**	**0.874**	**0.914**	**0.799**	**0.836**

TABLE IV

RESULTS OF *ContextNet* WITH AND WITHOUT EDGE LOSS. WITH THE EDGE LOSS, THE PERFORMANCE IS BETTER

	JuddDB		DUT-OMRON		THUR15K		SED2		ECSSD		Pascal-S	
	F_β	AUC	F_β	AUC	F_β	AUC	F_β	AUC	F_β	AUC	F_β	AUC
w/o Edge Loss	0.524	0.494	0.750	0.744	0.715	0.703	0.873	0.865	0.865	0.903	0.789	0.822
w/ Edge Loss	**0.534**	**0.508**	**0.762**	**0.770**	**0.721**	**0.717**	**0.877**	**0.883**	**0.874**	**0.914**	**0.799**	**0.836**

TABLE V

COMPARISON WITH FUSING FEATURES. OUR PROPOSED FUSING MAPS METHOD OUTPERFORMS METHOD WHICH FUSES FEATURES

	JuddDB		DUT-OMRON		THUR15K		SED2		ECSSD		Pascal-S	
	F_β	AUC	F_β	AUC	F_β	AUC	F_β	AUC	F_β	AUC	F_β	AUC
Fusing Features	0.520	0.486	0.734	0.724	0.704	0.686	0.873	0.871	0.855	0.887	0.776	0.805
Fusing Maps	**0.534**	**0.508**	**0.762**	**0.770**	**0.721**	**0.717**	**0.877**	**0.883**	**0.874**	**0.914**	**0.799**	**0.836**

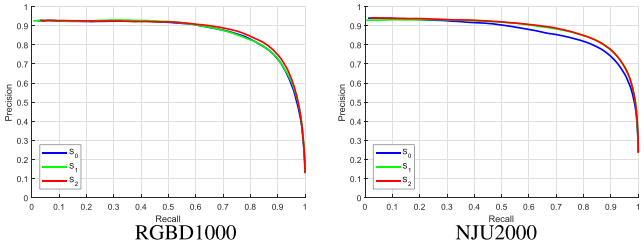


Fig. 13. Evaluate the effectiveness of depth refinement. Our depth refinement improves the performance mainly at the region with high recall, which is essential important for the final performance. Best viewed in color.

shows the result of *ContextNet* with fusing features and fusing maps. We can see that our method outperforms method which fuses features. This is benefited from the deep supervision in each branch which makes full use of features at different levels.

5) Depth Refinement: For the RGB-D saliency datasets, we evaluate the effectiveness of depth refinement. We show the comparison of PR-curves with and without depth refinement in Fig. 13. Experimental results show that the depth refinement improve the performance significantly especially in the region with high precision and high recall.

TABLE VI

PERFORMANCE AND SPEED COMPARISON WITH OTHER REGION-BASED CNN METHODS. OUR METHOD TAKES 0.4s FOR SEGMENTING IMAGE INTO REGIONS, AND ONLY 0.35s FOR NETWORK FORWARDING. OUR METHOD TAKES LESS TIME WHILE ACHIEVING BETTER PERFORMANCE

	F_β	AUC	Time (s)
RexNet [Ours]	0.893	0.937	0.40 + 0.35
MC [19]	0.822	0.852	1.63
LEGS [18]	0.827	0.855	2.27

6) Speed: We compare the speed with other region-based CNN methods. Our method is much faster since we deal with regions under end-to-end Fast R-CNN framework, while other region-based CNN methods forward network for each region. Table VI shows the comparison of performance and running time, the experiment is conduct on ECSSD dataset [26], it contains 1000 test images, we test on this dataset with a single NVIDIA GeForce GTX TITAN GPU and report the average time per image. We compare with MC [19] and LEGS [18] using the authors' public code. Our method takes 0.75s for each image, including 0.4s for segmenting image into regions using superpixel and edges and only 0.35s for network forwarding. Our method takes less time while achieving better performance.

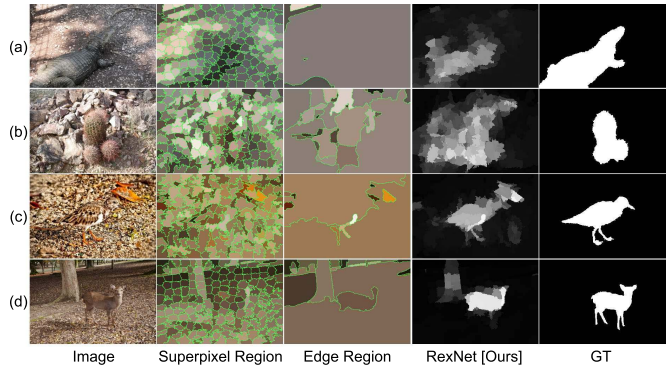


Fig. 14. Some failure cases of our method. These images are with extreme low-contrast scenes, which makes it difficult to segment into correct regions, thus influencing the final results. (a, b) both superpixel and edge segmentation fail, the result is bad. (c, d) the boundary between object and background is a bit clearer, thus the result is much better than (a) and (b).

E. Failure Cases

Our proposed framework achieves state-of-the-art performance. However, as the *RegionNet* is based on the segmentation of images, when the image is with extreme low contrast and the boundary between object and background is blurry, the segmentation may fail and thus influencing the final performance. Fig. 14 shows some failure examples. These images are all in scene with low contrast, when both superpixel and edge segmentation fail, the performance decreases much. Note that in Fig. 14 (c) and (d), though the scene is low-contrast, the boundary between object and background is a bit clearer, thus the result is much better than Fig. 14 (a) and (b).

VIII. Conclusion

In this paper, we propose *RexNet* which generates saliency map end-to-end and with sharp object boundaries. In the proposed framework, image is first segmented into two scales of complementary regions: superpixel regions and edge regions. The network then generates saliency score of regions end-to-end and context in multiple layers are considered to fuse with region saliency scores. The proposed *RexNet* achieves both clear detection boundary and multi-scale contextual robustness simultaneously for the first time, thus achieves an optimized performance. We also extend the proposed framework to RGB-D saliency detection by depth refinement. Experiments on benchmark RGB and RGB-D datasets demonstrate that the proposed method achieves state-of-the-art performance.

References

[1] Y. Wei *et al.* (2015). "STC: A simple to complex framework for weakly-supervised semantic segmentation." [Online]. Available: https://arxiv.org/abs/1509.03150

[2] V. Mahadevan and N. Vasconcelos, "Saliency-based discriminant tracking," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 1007–1013.

[3] S. Hong, T. You, S. Kwak, and B. Han, "Online tracking by learning discriminative saliency map with convolutional neural network," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 597–606.

[4] B. Lei, E.-L. Tan, S. Chen, D. Ni, and T. Wang, "Saliency-driven image classification method based on histogram mining and image score," *Pattern Recognit.*, vol. 48, no. 8, pp. 2567–2580, 2015.

[5] B. Li, W. Xiong, O. Wu, W. Hu, S. Maybank, and S. Yan, "Horror image recognition based on context-aware multi-instance learning," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5193–5205, Dec. 2015.

[6] 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.

[7] L. Zhang, M. H. Tong, T. K. Marks, H. Shan, and G. W. Cottrell, "SUN: A Bayesian framework for saliency using natural statistics," *J. Vis.*, vol. 8, no. 7, p. 32, Dec. 2008.

[8] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga, "Saliency estimation using a non-parametric low-level vision model," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 433–440.

[9] Y. Zhai and M. Shah, "Visual attention detection in video sequences using spatiotemporal cues," in *Proc. ACM MM*, 2006, pp. 815–824.

[10] M.-M. Cheng, G.-X. Zhang, N. J. Mitra, X. Huang, and S.-M. Hu, "Global contrast based salient region detection," in *Proc. CVPR*, 2011, pp. 569–582.

[11] A. Borji, M.-M. Cheng, H. Jiang, and J. Li. (2014). "Salient object detection: A survey." [Online]. Available: https://arxiv.org/abs/1411.5878

[12] T. Liu, J. Sun, N.-N. Zheng, X. Tang, and H.-Y. Shum, "Learning to detect a salient object," in *Proc. CVPR*, Jun. 2007, pp. 1–5.

[13] K. Shi, K. Wang, J. Lu, and L. Lin, "Pisa: Pixelwise image saliency by aggregating complementary appearance contrast measures with spatial priors," in *Proc. CVPR*, 2013, pp. 2115–2122.

[14] P. Jiang, H. Ling, J. Yu, and J. Peng, "Salient region detection by UFO: Uniqueness, focusness and objectness," in *Proc. ICCV*, 2013, pp. 1976–1983.

[15] Y. Wei, F. Wen, W. Zhu, and J. Sun, "Geodesic saliency using background priors," in *Proc. ECCV*, 2012, pp. 29–42.

[16] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. CVPR*, 2014, pp. 2814–2821.

[17] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform," in *Proc. CVPR*, 2014, pp. 883–890.

[18] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. CVPR*, 2015, pp. 3183–3192.

[19] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. CVPR*, 2015, pp. 1265–1274.

[20] T. Chen, L. Lin, L. Liu, X. Luo, and X. Li, "DISC: Deep image saliency computing via progressive representation learning," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 27, no. 6, pp. 1135–1149, Jun. 2016.

[21] X. Li *et al.*, "DeepSaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3919–3930, Aug. 2016.

[22] R. Girshick, "Fast R-CNN," in *Proc. ICCV*, 2015, pp. 1440–1448.

[23] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Proc. ECCV*, 2014, pp. 818–833.

[24] Y. Li, J. Yosinski, J. Clune, H. Lipson, and J. Hopcroft, "Convergent learning: Do different neural networks learn the same representations?" in *Proc. ICLR*, 2016, pp. 196–212.

[25] X. Li, Y. Li, C. Shen, A. Dick, and A. Van Den Hengel, "Contextual hypergraph modeling for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 3328–3335.

[26] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. CVPR*, 2013, pp. 1155–1162.

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

[28] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2015, pp. 3431–3440.

[29] R. Ju, L. Ge, W. Geng, T. Ren, and G. Wu, "Depth saliency based on anisotropic center-surround difference," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Oct. 2014, pp. 1115–1119.

[30] H. Peng, B. Li, W. Xiong, W. Hu, and R. Ji, "RGBD salient object detection: A benchmark and algorithms," in *Proc. Eur. Conf. Comput. Vis.*, 2014, pp. 92–109.

[31] J. Ren, X. Gong, L. Yu, W. Zhou, and M. Y. Yang, "Exploiting global priors for RGB-D saliency detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, Jun. 2015, pp. 25–32.

[32] D. Feng, N. Barnes, S. You, and C. McCarthy, "Local background enclosure for RGB-D salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 2343–2350.

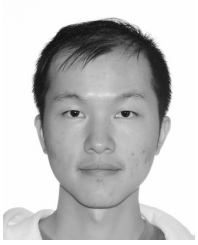
[33] B. Hariharan and P. Arbeláez, R. Girshick, and J. Malik, "Hypercolumns for object segmentation and fine-grained localization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 447–456.

[34] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 678–686.
[35] M. Liang and X. Hu, "Recurrent convolutional neural network for object recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2015, pp. 3367–3375.
[36] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. ICCV*, 2015, pp. 1395–1403.
[37] J. Zhang and S. Sclaroff, "Saliency detection: A Boolean map approach," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2013, pp. 153–160.
[38] L.-C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. L. Yuille, "Semantic image segmentation with deep convolutional nets and fully connected crfs," in *Proc. Int. Conf. Learn. Represent.*, 2015, pp. 1–2.
[39] X. Wang, H. Ma, and X. Chen, "Salient object detection via fast R-CNN and low-level cues," in *Proc. IEEE ICIP*, Sep. 2016, pp. 1042–1046.
[40] J. Dai, K. He, and J. Sun, "Instance-aware semantic segmentation via multi-task network cascades," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2016, pp. 3150–3158.
[41] P. Dollár and C. L. Zitnick, "Structured forests for fast edge detection," in *Proc. ICCV*, Dec. 2013, pp. 1841–1848.
[42] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "Slic superpixels compared to state-of-the-art superpixel methods," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 11, pp. 2274–2282, Nov. 2012.
[43] K. Simonyan and A. Zisserman. (2014). "Very deep convolutional networks for large-scale image recognition." [Online]. Available: https://arxiv.org/abs/1409.1556
[44] C.-Y. Lee, S. Xie, P. Gallagher, Z. Zhang, and Z. Tu, "Deeply-supervised nets," in *Proc. AISTATS*, vol. 2. 2015, p. 6.
[45] Y. Jia *et al.*, "Caffe: Convolutional architecture for fast feature embedding," in *Proc. ACM MM*, 2014, pp. 675–678.
[46] O. Russakovsky *et al.*, "ImageNet large scale visual recognition challenge," *Int. J. Comput. Vis.*, vol. 115, no. 3, pp. 211–252, Dec. 2015.
[47] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. CVPR*, 2013, pp. 3166–3173.
[48] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. CVPR*, 2009, pp. 1597–1604.
[49] A. Borji, "What is a salient object? A dataset and a baseline model for salient object detection," *IEEE Trans. Image Process.*, vol. 24, no. 2, pp. 742–756, Feb. 2015.
[50] S. Alpert, M. Galun, R. Basri, and A. Brandt, "Image segmentation by probabilistic bottom-up aggregation and cue integration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 315–327, Feb. 2012.
[51] M.-M. Cheng, N. J. Mitra, X. Huang, and S.-M. Hu, "SalientShape: Group saliency in image collections," *Vis. Comput.*, vol. 30, no. 4, pp. 443–453, 2014.
[52] 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.*, Jun. 2014, pp. 280–287.
[53] A. Borji, M.-M. Cheng, H. Jiang, and J. Li, "Salient object detection: A benchmark," *IEEE Trans. Image Process.*, vol. 24, no. 12, pp. 5706–5722, Dec. 2015.
[54] F. Perazzi, P. Krähenbühl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. CVPR*, Jun. 2012, pp. 733–740.
[55] 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.
[56] Z. Liu, W. Zou, and O. Le Meur, "Saliency tree: A novel saliency detection framework," *IEEE Trans. Image Process.*, vol. 23, no. 5, pp. 1937–1952, May 2014.
[57] H. Li, H. Lu, Z. Lin, X. Shen, and B. Price, "Inner and inter label propagation: Salient object detection in the wild," *IEEE Trans. Image Process.*, vol. 24, no. 10, pp. 3176–3186, Oct. 2015.
[58] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 fps," in *Proc. ICCV*, 2015, pp. 1404–1412.

Xiang Wang received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2014, where he is currently pursuing the Ph.D. degree. His research interests are computer vision and machine learning, with particular interests in salient object detection and semantic segmentation.

Huimin Ma (M'11) received the M.S. and Ph.D. degrees in mechanical electronic engineering from the Beijing Institute of Technology, Beijing, China, in 1998 and 2001, respectively. She is currently an Associate Professor with the Department of Electronic Engineering, Tsinghua University, and the Director of 3D Image Simulation Laboratory. She was a Visiting Scholar with University of Pittsburgh in 2011. She is also the Secretary-General of China Society of Image and Graphics. Her research and teaching interests include 3D object recognition and tracking, system modeling and simulation, and psychological base of image cognition.

Xiaozhi Chen received the B.S. degree in electronic engineering from Tsinghua University, Beijing, China, in 2012, where he is currently pursuing the Ph.D. degree. His research interests include computer vision especially on 3D object detection and machine learning.



Shaodi You received the bachelor's degree from Tsinghua University, China, in 2009, the. M.E. and Ph.D. degrees from The University of Tokyo, Japan, in 2015 and 2012. He is currently a Research Scientist with Data61-CSIRO (formerly known as NICTA), Australia. He also serves as an Adjunct Lecturer with Australian National University, Australia. His research interests are physics based vision, nonrigid 3D geometry and perception and learning based vision. He is currently the Chair of IEEE Computer Society, Australian Capital Territory Section, Australia. He is the Program Chair of ICCV2017 Joint Workshop on Physics Based Vision meets Deep Learning. He serves as a Reviewer for TPAMI, IJCV, TIP, CVPR, ICCV, and SIGGRAPH.