The Visual Computer
https://doi.org/10.1007/s00371-019-01778-4

ORIGINAL ARTICLE

Robust salient object detection for RGB images

Zhengyi Liu[1] · Qian Xiang[1] · Jiting Tang[1] · Yuan Wang[1] · Peng Zhao[1]

Abstract
Recent advances in supervised salient object detection modeling have resulted in significant performance improvements on benchmark dataset. However, most of the existing salient object detection models assume that an image contains at least one salient object. Such an assumption that often leads to their effectiveness may be impaired once they are applied to real-world scenes. To solve the problem, salient object existence prediction designed to judge whether the image contains salient objects is introduced into deep network to learn a better salient object detection model. For dense salient object detection task, high-level semantic feature is progressively hybrid upsampled from deep to shallow to remedy the spatial information loss guided by higher layer feature and saliency existence information. Our model is aware of non-salient image that contains no salient objects at all and thus reduces the false-positive rate. Experimental results show that our model wins similar multi-task model and outperforms state-of-the-art models in robustness and accuracy.

Keywords Salient object detection · Salient object existence prediction · Non-salient images · Robust · Multi-task learning

1 Introduction

Salient object detection mimics the ability of human visual system to detect the most attention-grabbing objects in a scene. It is pre-processor in image processing and has achieved remarkable improvement in recent years. But existing dataset mostly contains images with a single object or several objects in low clutter. They do not adequately reflect the complexity of images in the real world where scenes usually contain no salient objects at all or contain multiple objects amidst lots of clutter, which are called non-salient images. Current models for salient object detection have nearly saturated the performance over existing dataset but cannot achieve satisfactory performance on realistic scenes.

One way to solve this problem is to train the model on a dataset containing a large number of non-salient images, so that the model can learn how to detect non-salient objects directly from the data. As shown in Fig. 1, non-salient images contain no salient objects, for example, sky, grass, texture, densely distributed similar objects, fuzzy shape and region without semantics. Saliency maps generated by models (DSS [15], BMPM [58], R3Net [9], CPD [52] and BASNet [36]) directly are quite far from the ground truth, as shown in the first three lines within the blue box. Although the model can learn how to detect non-salient images through training, the addition of non-salient images will weaken the ability of model to detect salient objects, as shown in the last four lines within the red box. So the best approach is to design a model that can handle non-salient images.

Jiang et al. [19] propose a supervised learning approach for jointly addressing the salient object detection and existence prediction problems by the structural SVM framework to predict both image-level existence labels and pixel-level saliency values. Due to the use of manual extraction features rather than deep learning methods, there is still much room for improvement in the effectiveness of this approach. Zhang et al. [56] propose salient object subitizing (SOS) technique for improving the accuracy of salient object detection. If the

✉ Zhengyi Liu
liuzywen@ahu.edu.cn; 22927463@qq.com

Qian Xiang
xiangqianforth@qq.com

Jiting Tang
1796340141@qq.com

Yuan Wang
wangyuan.ahu@qq.com

Peng Zhao
18868519@qq.com

[1] Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei, China

Published online: 05 December 2019

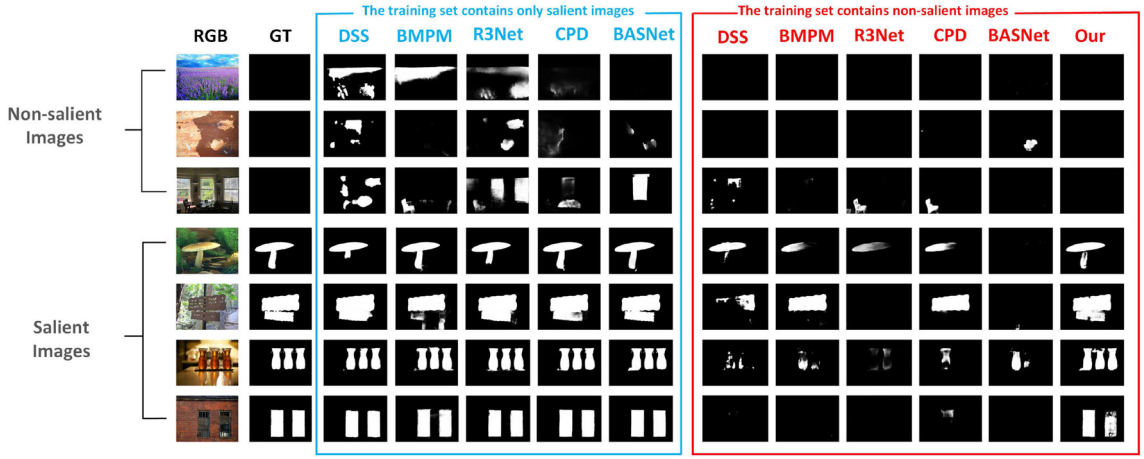


Fig. 1 Saliency maps of five state-of-the-art models and ours. The models in the blue box are trained only in the datasets containing salient images, while the models in the red box are trained in the datasets containing salient and non-salient images

image contains no salient objects, then an all-black saliency map is generated directly. Hou et al. [17] predict the saliency existence of the input image by introducing another branch into salient object detection network with short connections. It improves the accuracy of salient object existence prediction based on salient object detection network but does not consider jointly training both tasks.

This paper is the first to jointly train end-to-end back-propagation network for robust salient object detection and salient object existence prediction tasks. Both tasks are mutually beneficial from each other. Salient object existence prediction network provides strong guidance to salient object detection by reducing false positives and producing coherent saliency maps. Besides, salient object detection in turn helps salient object existence prediction achieve higher accuracy too. The robust model can detect salient objects directly instead of detecting salient objects after salient object existence prediction when facing non-salient images proposed by Zhang et al. [56].

Shared feature extractor generates effective features for both salient object detection and salient object existence prediction tasks. Deeper convolutional features are able to capture discriminative object parts and served as salient object existence prediction, but not enough for salient object detection task to highlight distinct boundaries. Then upsampling hybrid module is progressively performed from deep to shallow to refine salient objects step by step guided by salient object existence prediction. Salient object detection task and salient object existence prediction task are jointly trained end to end. Our key contributions are:

(1) Robust salient object detection model is proposed and shows its robustness especially in improving the accuracy of salient object detection by being aware of existence of salient objects and reducing false-positive detections on non-salient images.

(2) Joint training end-to-end network is used for salient object detection and salient object existence prediction. Robustness of salient object detection is enhanced by salient object existence prediction tasks. Accuracy of salient object existence prediction is promoted by salient object detection task. Both tasks benefit from each other.

2 Related work

2.1 Salient object detection

Saliency detection includes eye-fixation prediction (FP) [35,46,49] which can predict scene locations where a human observer may fixate and salient object detection (SOD) which can capture most visually conspicuous objects or regions. The initial research of salient object detection focuses on salient object detection in an image [18,32,33,41,59], then it is extended to RGB-D saliency detection [5,6,12,31,40,44,60] in which depth information can be utilized, co-saliency detection [24,42,51,57] in which inner and inter-saliency constraint need be considered simultaneously, and video saliency detection [13,37,39,47,48] in which temporal and spatial relationship are explored. Our work focuses on SOD in an image.

Over the past two decades, a large number of salient object detection methods have been developed. The majority of existing methods are based on hand-crafted features. Li et al. [25] present a method that involves the transfer of annotations from an example image to an input image and a coarse-to-fine saliency optimization framework. Recent years, methods based on deep learning have demonstrated impressive performance. For example, Hou et al. [15,17] introduce dense short connections into the skip layers within the holistically nested edge detection (HED) architecture [53] to get rich multi-scale

features for salient object detection. Zhang et al. [58] propose a controlled bi-directional passing of features between shallow and deep layers to obtain accurate predictions. Deng et al. [9] develop a recurrent residual refinement network for saliency maps refinement by incorporating shallow and deep layers features alternately. Qin et al. [36] propose a end-to-end predict-refine architecture BASNet. It is able to capture both large-scale and fine structures, e.g., thin regions, holes, and produce salient object detection maps with clear boundaries. Wu et al. [52] propose a cascaded partial decoder framework, which discards low-level features to reduce the complexity of deep aggregation models and utilizes generated relatively precise attention map to refine high-level features to improve the performance. Li et al. [27] employ a multi-scale cascade structure and a refinement module to filter out errors. It better consolidates contextual information and intermediate saliency priors.

Although aforementioned approaches employ powerful CNNs and make remarkable success in salient object detection, they produce unsatisfactory results in dealing with non-salient images problems as shown in Fig. 1. As a result, there is still a large room for performance improvements.

2.2 Salient object existence prediction

Wang et al. [45] exploit hand-crafted global features from multiple saliency information to directly predict the existence and the position of the salient object in web images by random forest. The purpose of this work is different from ours and focuses more on location of salient object whose result is expressed by bounding box enclosing the salient object region. Zhang et al. [56] investigate not only existence but also counting the number of salient objects based on holistic cues. If the image contains no salient objects, then an all-black saliency map is generated directly while salient object detection is not performed. Jiang et al. [19] propose a supervised learning approach for jointly addressing the salient object detection and existence prediction problems by the structural SVM framework to predict both image-level existence labels and pixel-level saliency values. Hou et al. [17] predict the saliency existence of the input image by introducing another branch into salient object detection network with short connections. It does not consider jointly training of both tasks, but only improve the accuracy of salient object existence prediction based on salient object detection network.

In this paper, we focus on recognizing saliency existence and locating salient objects. By incorporating image-level label, better performance of salient object detection with pixel level can be achieved.

2.3 Multi-task models

Salient object detection is to identify the most visually distinctive objects or regions in an image and then segment them out from the background. Semantic segmentation, image classification, salient object contour detection, subitizing and salient object existence prediction are discussed to guide salient object detection in recent years. Li et al. [26] set up a multi-mask learning scheme for exploring the intrinsic correlations between saliency detection and image semantic segmentation. Cholakkal et al. [8] propose a framework for top-down salient object detection that incorporates a tightly coupled image classification module. Wang et al. [43] propose to use image-level tags as weak supervision to learn to predict pixel-level saliency maps solving the problem that train DNNs require costly pixel-level annotations. Li et al. [23] propose to use the combination of a coarse salient object activation map from the classification network and saliency maps generated from unsupervised methods as pixel-level annotation, to train fully convolutional networks for salient object detection supervised by these noisy annotations. Wang et al. [22] design a deep multi-scale refinement network for both salient region detection and salient object contour detection. Zhuge et al. [61] propose a fully convolutional networks to integrate multi-level convolutional features recurrently with the guidance of object boundary information. He et al. [14] detect salient objects with the aid of subitizing. Jiang et al. [19] propose to jointly train the salient object detection and existence prediction problems by the structural SVM framework. Li et al. [28] graft salient object detection decoder onto the existing contour detection network to form a multi-task network architecture without using any manually labeled salient object masks. Although it implements joint training of two tasks, both tasks belong to the pixel-level segmentation task, lack of more multi-modal information. Hou et al. [16] aim at solving pixel-wise binary problems, including salient object detection, skeleton extraction and edge detection, by introducing a horizontal cascade encoder architecture. But this general structure cannot handle multiple tasks at the same time, and does not consider the complementarity between multiple tasks. Wang et al. [50] design a neural network that has two branches for attention box prediction (ABP) and aesthetics assessment (AA) to crop photograph with the best aesthetic quality. ABP subnetwork is responsible for inferring the initial cropping, and the AA network determines the final cropping. ABP task is followed by AA task. These two tasks are not learned simultaneously. They only share several convolutional blocks in the bottom of network.

CNN model trained in end-to-end manner for both salient object detection and salient object existence prediction tasks is unexplored in aforementioned literature.

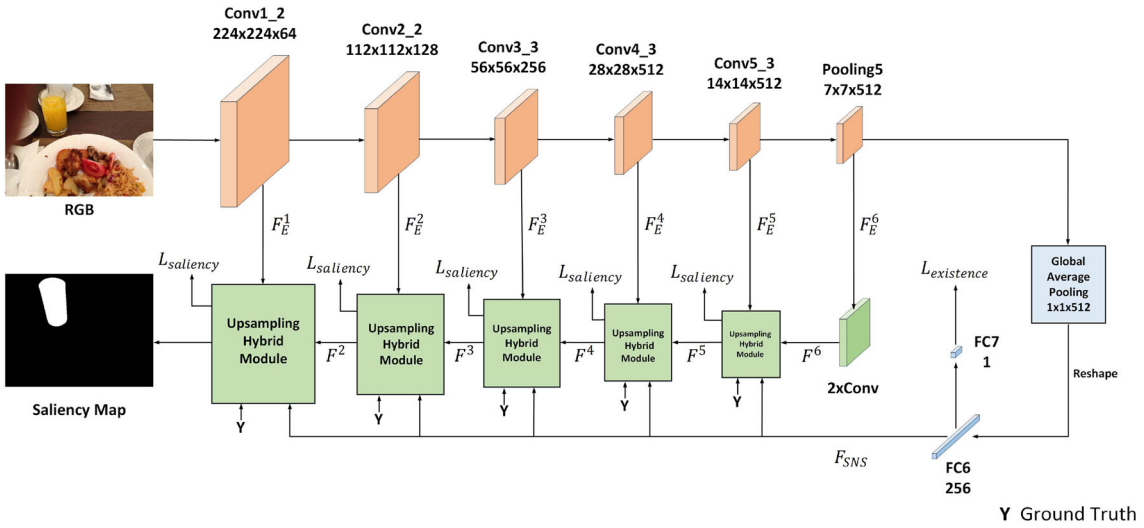


Fig. 2 The overall framework of our model

3 The proposed method

3.1 Overview

The proposed method is a multi-task deep convolutional neural network model, which includes salient object existence prediction and salient object detection. Salient object existence prediction only judges whether the image contains salient objects, so it only needs high-level semantic information. Salient object detection needs to achieve dense saliency map, but high-level semantic information loses the spatial information, so high-level semantic information needs to be transmitted from deep to shallow and further integrated with high-resolution low-level cues to produce high precision saliency map with distinct boundaries.

Figure 2 overviews the network architecture. Firstly, all input images are down-sampled to fixed resolution 256 * 256. Then shared feature extractor uses the VGG-16 net [38] that removes the fully connected layer to extract shared multi-level side-output features $F_E^m (m = 1, \ldots, 6)$ of the input images I.

$$F_E^m = \begin{cases} \tau(Conv(\tau(Conv(I)))), & m = 1 \\ \tau(Conv(\tau(Conv(Maxpooling(F_E^1))))), & m = 2 \\ \tau(Conv(\tau(Conv(\tau(Conv(Maxpooling(F_E^{m-1}))))))), & m = 3, 4, 5 \\ Maxpooling(F_E^5), & m = 6 \end{cases} \quad (1)$$

where $Conv(\cdot)$ denotes convolution operation with 3×3 kernel size, $\tau(\cdot)$ denotes ReLU activation function operation and $MaxPooling(\cdot)$ denotes a max pooling operation.

Then a global average pooling layer is used to reduce the dimension of the feature for salient object existence prediction task. Next the global feature is reshaped and fed into custom two fully connected layers to determine whether the image contains salient objects. Besides, salient object detection network consists of five different level upsampling hybrid modules. Each module aggregates the deconvolution features from deeper layer, the encoder features from current layer of VGG-16 net, and the image level features from salient object existence prediction network to progressively refine saliency map. Salient object detection task and salient object existence prediction task are jointly trained end to end.

3.2 Salient object existence prediction network

At present many classification algorithms based on neural network almost always add three fully connected layers with 4096, 4096 and 1 neurons after the convolution layers to conduct the characteristic vectorization. In [17], three fully connected layers with 1024, 1024 and 2 neurons are adopted to reduce training cost. Even so the running time of the network is very long for a great number of parameters need to be trained.

In the paper the dimension of output feature F_E^6 of VGG-16 net is $7 * 7 * 512$; then the global average pooling is adopted to accelerate dimension reduction further to generate $1 * 1 * 512$ global feature. And then global feature is reshaped and subsequently fed into custom two fully connected layers FC6, FC7 with 256,1 neurons, respectively, for salient object existence prediction. The selection of the number of neurons in custom fully connected layers takes both network training cost and guidance on salient object detection task into account. It first improves training speed by reducing the parameters and at the time aggregates with convolutional feature from shared feature extractor more easily for salient

object boundary optimization. The detail can be expressed by:

$$F_{SNS} = \tau(FC(Reshape(GAP(F_E^6)), 256)) \quad (2)$$

$$\hat{z} = Sigmoid(FC(F_{SNS}, 1)) \quad (3)$$

where $GAP(\cdot)$ denotes global average pooling, $Reshape(\cdot)$ denotes the dimension change operation, $FC(x, n)$ denotes the operation of performing a fully connected layer on the input feature x and the number of neurons output by the fully connected layer is n, $Sigmoid(\cdot)$ denotes sigmoid activation function.

3.3 Salient object detection network

The side-output $F_E^m (m = 1, \ldots, 6)$ of VGG-16 net expresses different level feature of input image I. Deep features can coarsely detect and localize salient objects, but loss the boundaries and subtle structures due to multiple pooling operations. Salient object detection network begins with deep feature F_E^6, and then performs two $3*3$ convolution operation for reducing the dimension of the feature to form the feature F^6, and then performs five successively progressive upsampling hybrid modules, and generates saliency map finally.

The structure of each upsampling hybrid module is shown in Fig. 3. Its detailed information is shown in Table 1. The feature $F^{m+1}(m = 1, \ldots, 5)$ from higher layer module output is transformed to F_D^m which doubles the resolution of the feature F^{m+1} by deconvolutional operation. The feature F_{SNS} from FC6 layer of salient object existence prediction network implies whether the image contains salient objects.

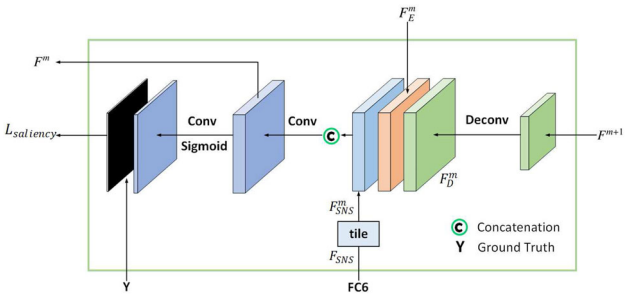


Fig. 3 The architecture of upsampling hybrid modules

It is tiled by replicating multiple times to the same size with side-output feature F_E^m, respectively, for concatenating easily, and it is defined as F_{SNS}^m.

The existence feature F_{SNS}, side-output feature F_E^m and feature F^{m+1} from higher layer module are fused together by one 1×1 convolutional layer to learn cooperative representations F^m:

$$F^m = \begin{cases} \tau(Conv_{1\times1}(D(F^{m+1}) \oplus tile(F_{SNS}) \oplus F_E^m)), & m = 1, 2, \ldots, 5 \\ \tau(Conv(\tau(Conv(F_E^m)))), & m = 6 \end{cases} \quad (4)$$

where $Conv_{1\times1}(\cdot)$ denotes convolution operation with 1×1 kernel size, D denotes deconvolution operation, $\oplus$ denotes concatenation operation, and $tile(\cdot)$ function denotes duplication operation.

The feature F^m is transmitted to next upsampling hybrid module and at the same time is performed by convolution operation with $1 * 1$ kernel again to generate side-output saliency map $\hat{Y}$ by sigmoid activation function.

$$\hat{Y}^m = Sigmoid(Conv_{1\times1}(F^m)) \quad (5)$$

where $\hat{Y}^m$ denotes side-output saliency map from the m^{th} upsampling hybrid model.

3.4 Multi-task learning

For an input image I, Y denotes the ground truth saliency map, z denotes the salient object existence prediction ground truth label indicating whether the image contains salient objects. Inspired by skip-connect structure [17], the loss $L_{\text{detection}}$ of salient object detection task is computed in all the upsampling hybrid modules and is defined as:

$$L_{\text{detection}} = \sum_{m=1}^{5} \mu^m * \phi(\hat{Y}^m, Y) \quad (6)$$

where $Y = \{y_i | i = 1, \ldots, N\}$ is pixel level of ground truth saliency map, $\hat{Y}^m = \{\hat{y}_i^m | i = 1, \ldots, N, m = 1, \ldots, 5\}$ is the side-output saliency maps from five upsampling hybrid modules, respectively, N is the number of pixels of input image, μ^m denotes the weight for the loss in the m-th module, and is set as 1 in the experiment, and $\phi(\cdot, \cdot)$ denotes the

Table 1 The size of each feature in the five upsampling hybrid modules. $k \times k \times c$ means $k \times k$ size and c channels

m	F^{m+1}	F_D^m	F_E^m	F_{SNS}^m	F^m
1	128 * 128 * 256	256 * 256 * 256	256 * 224 * 64	256 * 256 * 256	256 * 256 * 256
2	64 * 64 * 256	128 * 128 * 256	128 * 128 * 128	128 * 128 * 256	128 * 128 * 256
3	32 * 32 * 256	64 * 64 * 256	64 * 64 * 256	64 * 64 * 256	64 * 64 * 256
4	16 * 16 * 256	32 * 32 * 256	32 * 32 * 512	32 * 32 * 256	32 * 32 * 256
5	8 * 8 * 256	16 * 16 * 256	16 * 16 * 512	16 * 16 * 256	16 * 16 * 256

distance between the side-output saliency map $\hat{Y}^m$ and the ground truth map Y, and is defined as:

$$\phi(\hat{Y}^m, Y) = \sum_{i=1}^{N} y_i \log \hat{y}_i^m + (1 - y_i) \log(1 - \hat{y}_i^m) \tag{7}$$

The loss $L_{\text{existence}}$ of salient object existence prediction task between the actual output $\hat{z}$ and ground truth z is defined as:

$$L_{\text{existence}} = \varphi(\hat{z}, z) = z \log \hat{z} + (1 - z) \log(1 - \hat{z}) \tag{8}$$

where $\varphi(.,.)$ represents the cross-entropy.

We perform multi-task learning by taking the salient object detection task and the salient object existence prediction task into account. Therefore, the final loss function L_{final} is given by:

$$L_{\text{final}} = \alpha * L_{\text{detection}} + \beta * L_{\text{existence}} \tag{9}$$

where α and β are the weight for the loss $L_{\text{detection}}$ of salient object detection task and the loss $L_{\text{existence}}$ of salient object existence prediction task and is set as 10:1.

Network is jointly trained for both salient object detection and salient object existence prediction tasks. Restricted by final loss function L_{final}, robustness of salient object detection is enhanced by salient object existence prediction tasks, and at the same time accuracy of salient object existence prediction is promoted by salient object detection task. The algorithm for the model is shown in Algorithm 1.

Algorithm 1: The Algorithm for Our Network Model

Input: I is a RGB image with 256*256 size, Y is a ground truth image with 256*256 size, z is a image-level label;
Output: $\hat{Y}^m$ is the m^{th} side-output saliency map with 256*256 size, $\hat{z}$ is a corresponding image-level predict label;

$L_{final} = +\infty$;
while $L_{final} > 0$ **do**
  $F_E^1 = \tau(Conv(\tau(Conv(I))))$;
  $F_E^2 = \tau(Conv(\tau(Conv(Maxpooling(F_E^1)))))$;
  for $m = 3$ *to* 5 **do**
    $F_E^m = \tau(Conv(\tau(Conv(\tau(Conv(Maxpooling(F_E^{m-1}))))))))$;
  $F_E^6 = Maxpooling(F_E^5)$;
  $F_{SNS} = \tau(FC(Reshape(GAP(F_E^6)), 256))$;
  $\hat{z} = Sigmoid(FC(F_{SNS}, 1))$;
  $L_{existence} = \varphi(\hat{z}, z)$;
  $F^6 = \tau(Conv(\tau(Conv(F_E^6))))$;
  $L_{detection} = 0$;
  for $m = 5$ *to* 1 **do**
    $F^m = Conv_{1\times1}(D(F^{m+1}) \bigoplus tile(F_{SNS}) \bigoplus F_E^m)$;
    $\hat{Y}^m = Sigmoid(Conv_{1\times1}(F^m))$;
    $L_{detection} = L_{detection} + \phi(\hat{Y}^m, Y^m)$;
  $L_{final} = \alpha * L_{detection} + \beta * L_{existence}$

4 Experiments

In this section, we introduce dataset and evaluation criteria and report the performance of our robust salient object detection model.

4.1 Dataset

Existing dataset [3,4,7,21,29,30,34,54,55] assumes that an image contains at least one salient object and thus discards images that do not contain salient objects. Fan et al. [11] propose a dataset SOC [11], which includes 3000 salient images and 3000 non-salient images from daily object categories, making them closer to real-world scenarios and more challenging than existing dataset. All the images are divided into 3600 images for training, 1200 images for validating and 1200 images for testing. JSOD [19] contains 6182 background images, 6237 images from MSRA10K [7], and 2419 images for testing. DUT-OMRON [55] consists of 5168 high-quality images that have one or more salient objects. HKU-IS [21] consists of 4447 images containing multiple salient objects with low color contrast or overlapping with the image boundary. MSRA-B [30] contains 5000 images from hundreds of different categories. ECSSD [54] contains 1000 semantically meaningful but structurally complex natural images.

SOC [11] training set which contains 1800 images with salient objects and 1800 images without salient objects is selected as the training set for all experiments. In addition, all experiments find the optimal solution on the SOC validation set which contains 1200 images, and then test on other datasets.

4.2 Evaluation metrics

We evaluate the performance of our model as well as other state-of-the-art salient object detection models using four metrics, including precision–recall (PR) curves, F-measure, mean absolute error (MAE) and S-measure. The precision value is the ratio of ground truth salient pixels in the predicted salient region, and the recall value is defined as the percentage of the detected salient pixels and all ground truth area. The precision and recall are calculated by thresholding the predicted saliency map and comparing it with the corresponding ground truth. Taking the average of precision and recall of all images in the dataset, we can plot the precision-recall curve at different thresholds. The F-measure is an overall performance indicator; it is computed by the weighted harmonic of precision and recall:

Table 2 The comparison of our model with SSVM [19]

Models	AP↑		F_β ↑		MAE↓		
	MSRA-B	ECSSD	MSRA-B	ECSSD	MSRA-B	ECSSD	JSOD
SSVM [19]	0.8830	0.7010	**0.8160**	0.6890	0.1070	0.2140	0.0510
Our	**0.9143**	**0.9364**	0.8032	**0.7938**	**0.0740**	**0.0805**	**0.0408**

$$F_\beta = \frac{(1+\beta^2)\cdot \text{Precision}\cdot \text{Recall}}{\beta^2\cdot \text{Precision} + \text{Recall}} \tag{10}$$

where β^2 is set to 0.3 to weight precision more than recall as suggested in [2]. Besides PR curve and F-measure, we also calculate the mean absolute error (MAE) to measure the average difference between predicted saliency map and ground truth. It is computed as:

$$\text{MAE} = \frac{1}{W\times H}\cdot\sum_{x=1}^{W}\sum_{y=1}^{H}|S(x,y)-G(x,y)| \tag{11}$$

where S and G are predicted saliency map and ground truth, respectively.

Due to the lack of global information consideration, above three evaluation indexes only consider the error of a single pixel, which leads to inaccurate evaluation. Therefore, a new evaluation metric recently proposed is also adopted. Metric S-measure [10] solves the problem of structural measurement from the perspective of region-aware (S_λ) and object-aware (S_o) and is defined as:

$$S_\lambda = \lambda * S_o + (1-\lambda) * S_\gamma \tag{12}$$

where $\lambda \in [0, 1]$ is the balance parameter, S_o and S_λ are the object-aware and region-aware structural similarity, respectively. We set $\lambda = 0.5$ as suggested in [10]. At last, in order to compare with SSVM [19], we increase the average precision (AP) score defined as the area under the PR curve.

4.3 Implementation details

The proposed model is implemented in Python with Tensorflow toolbox [1]. It is evaluated on a machine equipped with one GTX 1080Ti GPUs (with 11G memory). It is fine-tuned from an initialization with the pre-trained VGG-16 net model based on aforementioned dataset. In the experiments we minimize objective function using Adaptive Moment Estimation (Adam) [20], with a batch size of 1, and the learning rate is initialized as 0.00001. No data augmentation operations are used during the experiment. The training process of the model takes about 12 hours and converges after 60 epochs. During testing, the model runs about 20 fps with 256×256 resolution.

4.4 Performance comparison with SSVM

SSVM [19] is similar to ours for the same multi-task learning about salient object detection and salient object existence prediction. Unfortunately SSVM [19] provides no code but dataset JSOD only, so results in the paper [19] are directly compared with ours.

Our model is trained on the training set of SOC dataset. The same testing set as SSVM [19] are adopted in comparison process. They are MSRA-B [30], ECSSD [54] and testing set of JSOD [19]. Comparison results are shown in Table 2, the best result of each column is highlighted in bold, ↑ denotes the bigger value the better, and ↓ denotes the smaller value the better. For most metrics, our model achieves better performance than SSVM [19]. This also demonstrates the robustness of our model which can perform well in different testing set.

4.5 Performance comparison with state-of-the-art models in robustness and accuracy

The robustness of the model depends on performance on testing dataset. Since the training set used by existing models contains almost no non-salient images, direct comparisons are unfair. So we consider that all experiments should be retrained using the same training set. Five representative methods in which the authors provide the source codes are selected for the comparison. The five methods are DSS17 [15], BMPM18 [58], R3Net18 [9], CPD19 [36] and BASNet19 [52]. They are retrained using the same training set as ours to achieve fair comparison and reduce the workload. None of the experiments uses data augmentation operations.

Table 3 and Fig. 4 demonstrate our method outperforms the comparison methods across the aforementioned public datasets in terms of S-measure, F-measure and MAE evaluation metrics. Since the existing models are not designed to take into account the influence of non-salient images on the models, the ability of the model to detect salient objects will decrease after adding some non-salient images in the training process. The method of this paper combines the image-level classification features and the pixel-level semantic features, and jointly trains the loss of two tasks. The experimental results fully prove the robustness of the model. Figure 1 shows the visual comparison of saliency maps produced by

Table 3 Quantitative comparisons on six datasets

Models	SOC			JSOD			DUT-OMRON			HKU-IS			MSRA-B			ECSSD		
	S-measure↑	F_β ↑	MAE↓	S-measure↑	F_β ↑	MAE↓	S-measure↑	F_β ↑	MAE↓	S-measure↑	F_β ↑	MAE↓	S-measure↑	F_β ↑	MAE↓	S-measure↑	F_β ↑	MAE↓
DSS17	0.8833	0.3599	0.0466	0.8788	0.3939	0.0486	0.6798	0.5950	0.0949	0.7983	0.8073	0.0671	0.7717	0.7678	0.0888	0.7566	0.7640	0.1024
BMPM18	0.8865	0.3646	0.0451	0.8762	0.3905	0.0490	0.6873	0.6031	0.0917	0.7848	0.7910	0.0710	0.7701	0.7655	0.0877	0.7462	0.7471	0.1059
R3Net18	0.8835	0.3556	0.0520	0.8655	0.3893	0.0577	0.6321	0.5172	0.1108	0.7601	0.7766	0.0871	0.7514	0.7588	0.1038	0.7609	0.7702	0.1028
CPD19	0.8791	0.3572	0.0490	0.8568	0.3676	0.0556	0.6622	0.5783	0.0986	0.7725	0.7833	0.0765	0.7344	0.7262	0.1013	0.7202	0.7183	0.1153
BASNet19	0.8305	0.2920	0.0679	0.8520	0.3419	0.0568	0.6470	0.5222	0.1003	0.6805	0.6318	0.1080	0.7065	0.6546	0.1085	0.6260	0.5513	0.1508
Our	0.8921	0.3657	0.0427	0.8994	0.4103	0.0408	0.7188	0.6497	0.0819	0.8221	0.8274	0.0585	0.8106	0.8032	0.0740	0.7938	0.7938	0.0845

some state-of-the-art methods and our method. The first three lines are about non-salient images, and the last four lines are about salient images. The saliency maps in the blue box are directly testing results of five methods. The saliency maps in the red box are testing results after retaining the models based on SOC training dataset which is the same as ours. From the visual performance comparison, we can see that the direct testing results of five models in non-salient images are not good, after retraining, they can achieve the better performance, but the effect of detecting salient object declines. In contrast, our method is consistent good at both salient images and non-salient images. Table 4 shows the test runtime of different methods. As can be seen, there is no advantage in the speed of our method due to multi-task learning. But it is superior to the others in the performance, which can be seen in Table 3.

4.6 Ablation study

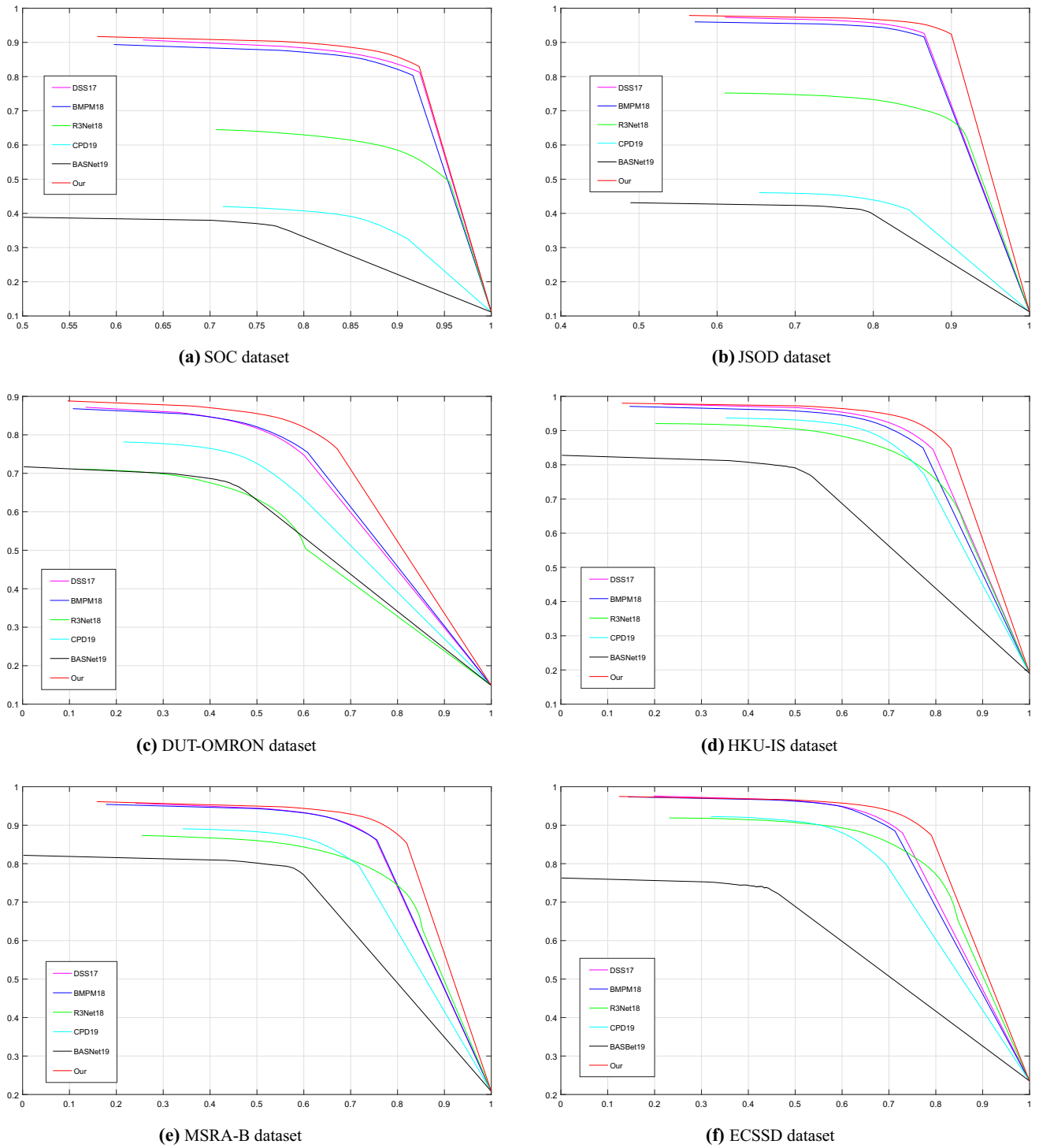
At first we verify the effectiveness of salient object existence prediction branch. The experiment is named Our, which removes the salient object existence prediction branch in our method, and the image-level features from $fc7$ are not added to the upsampling hybrid module. From Table 5, we can see that salient object existence prediction branch plays an important role in saliency detection.

Then the classic model DVA [46] in the field of salient object detection is selected as the experimental model. We add the salient object existence prediction branch and three upsampling hybrid module to DVA, the new experiment named DVA+. From Table 6, we can see that DVA+ equipped with the salient object existence prediction branch shows better performance than DVA. Therefore, the results verify our first contribution.

At last, we verify whether salient object detection branch provides guidance to salient object existence prediction. The experiment is named single-task existence prediction, which contains VGG-16 network and two custom fully connected layers about 256 and 1 neurons. From Table 7, we can see that our multi-task learning provides better guidance to salient object existence prediction and wins single-task pure existence prediction model with the same fully connected layers in classification accuracy. Therefore, the result verifies our second contribution.

4.7 Failure cases

The proposed method has a good detection effect in most cases. Due to complex image, failure cases are generated, as shown in Fig. 5. In the first line, some images are regarded as non-salient images and no salient objects are detected. In the second and third lines, some salient objects are missed or augmented. If the number of salient objects is specified,



(a) SOC dataset

(b) JSOD dataset

(c) DUT-OMRON dataset

(d) HKU-IS dataset

(e) MSRA-B dataset

(f) ECSSD dataset

Fig. 4 The PR curves of our model and other state-of-the-art models

Table 4 The runtime comparison among different models

Models	DSS17	BMPM18	R3Net18	CPD19	BASNet19	Our
Runtime (s)	0.0312	0.0273	0.0468	0.0234	0.0390	0.0508
Framework	Tensorflow	Tensorflow	Pytorch	Pytorch	Pytorch	Tensorlow

Table 5 Ablation study about the role of salient object existence prediction branch in our method

Models	SOC			JSOD			DUT-OMRON			HKU-IS			MSRA-B			ECSSD		
	S-measure↑	F_β ↑	MAE↓	S-measure↑	F_β ↑	MAE↓	S-measure↑	F_β ↑	MAE↓	S-measure↑	F_β ↑	MAE↓	S-measure↑	F_β ↑	MAE↓	S-measure↑	F_β ↑	MAE↓
Our	0.8876	0.3591	0.0457	0.8846	0.3896	0.0455	0.7004	0.5996	0.0895	0.8104	0.7990	0.0630	0.7797	0.7559	0.0844	0.7630	0.7516	0.1005
Our	0.8921	0.3657	0.0427	0.8994	0.4103	0.0408	0.7188	0.6497	0.0819	0.8221	0.8274	0.0585	0.8106	0.8032	0.0740	0.7938	0.7938	0.0845

Table 6 Ablation study about the role of salient object existence prediction branch in DVA [46]

Models	SOC			JSOD			DUT-OMRON			HKU-IS			MSRA-B			ECSSD		
	S-measure↑	F_β ↑	MAE↓	S-measure↑	F_β ↑	MAE↓	S-measure↑	F_β ↑	MAE↓	S-measure↑	F_β ↑	MAE↓	S-measure↑	F_β ↑	MAE↓	S-measure↑	F_β ↑	MAE↓
DVA	0.8534	0.3366	0.0599	0.8222	0.3420	0.0699	0.5984	0.4984	0.1146	0.7015	0.7295	0.1007	0.6642	0.6736	0.1272	0.6446	0.6614	0.1465
DVA+	0.8708	0.3450	0.0512	0.8521	0.3627	0.0578	0.6317	0.5338	0.1055	0.7433	0.7403	0.0838	0.7202	0.7100	0.1061	0.7014	0.6963	0.1184

Table 7 Accuracy comparison between single-task existence prediction network with our multi-task classification network

Models	Accuracy↑					
	SOC	JSOD	DUT-OMRON	HKU-IS	MSRA-B	ECSSD
Single-task existence prediction	0.9750	0.9016	0.7909	0.8588	0.8104	0.7580
Our	0.9800	0.9169	0.8101	0.8819	0.8390	0.8190

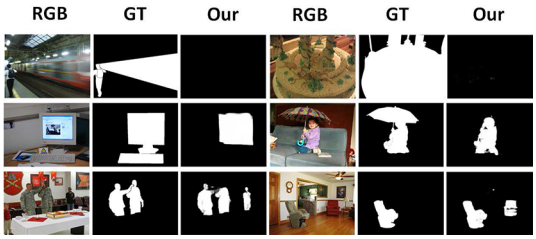


Fig. 5 Failure cases of our model

they will be not happened. So further research about salient object subitizing rather than saliency existence may improve the performance.

5 Conclusion

In this paper, we research the robustness problem of salient object detection model, which depends on performance in non-salient images. By combining salient object detection and salient object existence prediction tasks, robust salient object detection model trained with SOC dataset is proposed. It first achieves good performance and wins SSVM which is most similar to our training tasks. It then outperforms state-of-the-art models retrained in SOC dataset including non-salient images and shows its perception of non-salient images and robustness. It next verifies salient object existence prediction influence on salient object detection. At last salient object existence prediction with and without salient object detection guidance is contrasted. Results show that salient object detection branch plays a role in salient object existence prediction. Our work combines the image-level classification features and the pixel-level semantic features, and jointly trains the loss of two tasks. It can be applied in other visual application, for example, image retrieval and segmentation in which the class of object can be recognized and the region of object are segmented. Future research about salient object subitizing will improve the performance.

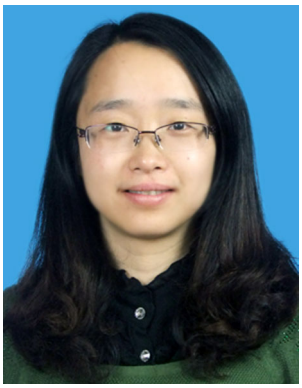
Acknowledgements We thank Ming-Ming Cheng from Nankai University for providing JSOD dataset. We also thank all anonymous reviewers for their valuable comments. This research is supported by National Natural Science Foundation of China (61602004), Natural Science Foundation of Anhui Province (1908085MF182) and University Natural Science Research Project of Anhui Province (KJ2019A0034).

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: a system for large-scale machine learning. OSDI **16**, 265–283 (2016)
2. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: IEEE Conference on Computer Vision and Pattern Recognition, 2009, CVPR 2009. IEEE, pp. 1597–1604 (2009)
3. Alpert, S., Galun, M., Basri, R., Brandt, A.: Image segmentation by probabilistic bottom-up aggregation and cue integration. In: IEEE Conference on Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE, pp. 1–8 (2007)
4. Borji, A., Cheng, M.M., Jiang, H., Li, J.: Salient object detection: a benchmark. IEEE Trans. Image Process. **24**(12), 5706–5722 (2015)
5. Chen, H., Li, Y.: Three-stream attention-aware network for rgb-d salient object detection. IEEE Trans. Image Process. **28**(6), 2825–2835 (2019)
6. Chen, H., Li, Y., Su, D.: Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for rgb-d salient object detection. Pattern Recognit. **86**, 376–385 (2019)
7. Cheng, M.M., Mitra, N.J., Huang, X., Torr, P.H., Hu, S.M.: Global contrast based salient region detection. IEEE Trans. Pattern Anal. Mach. Intell. **37**(3), 569–582 (2015)
8. Cholakkal, H., Johnson, J., Rajan, D.: A classifier-guided approach for top-down salient object detection. Signal Process. Image Commun. **45**(C), 24–40 (2016)
9. Deng, Z., Hu, X., Zhu, L., Xu, X., Qin, J., Han, G., Heng, P.A.: R3net: Recurrent residual refinement network for saliency detection. In: Proceedings of the 27th International Joint Conference on Artificial Intelligence, AAAI Press, pp. 684–690 (2018)
10. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: a new way to evaluate foreground maps. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4558–4567 (2017)
11. Fan, D.P., Cheng, M.M., Liu, J.J., Gao, S.H., Hou, Q., Borji, A.: Salient objects in clutter: bringing salient object detection to the foreground. arXiv preprint arXiv:1803.06091 (2018)
12. Fan, D.P., Lin, Z., Zhao, J.X., Liu, Y., Zhang, Z., Hou, Q., Zhu, M., Cheng, M.M.: Rethinking rgb-d salient object detection: models, datasets, and large-scale benchmarks. arXiv preprint arXiv:1907.06781 (2019)
13. Fan, D.P., Wang, W., Cheng, M.M., Shen, J.: Shifting more attention to video salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 8554–8564 (2019)
14. He, S., Jiao, J., Zhang, X., Han, G., Lau, R.W.: Delving into salient object subitizing and detection. In: 2017 IEEE International Conference on Computer Vision (ICCV), IEEE, pp. 1059–1067 (2017)
15. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 5300–5309 (2017)
16. Hou, Q., Liu, J., Cheng, M.M., Borji, A., Torr, P.H.: Three birds one stone: a unified framework for salient object segmentation, edge detection and skeleton extraction. arXiv preprint arXiv:1803.09860 (2018)
17. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.: Deeply supervised salient object detection with short connections. In: IEEE TPAMI (2019). https://doi.org/10.1109/TPAMI.2018.2815688

18. Huang, K., Gao, S.: Image saliency detection via multi-scale iterative CNN. In: The Visual Computer, pp. 1–13 (2019)
19. Jiang, H., Cheng, M.M., Li, S.J., Borji, A., Wang, J.: Joint salient object detection and existence prediction. Front. Comput. Sci. **13**(4), 778–788 (2017)
20. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. arXiv preprint arXiv:1412.6980 (2014)
21. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5455–5463 (2015)
22. Li, G., Xie, Y., Lin, L., Yu, Y.: Instance-level salient object segmentation. In: 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, pp. 247–256 (2017)
23. Li, G., Xie, Y., Lin, L.: Weakly supervised salient object detection using image labels. arXiv preprint arXiv:1803.06503 (2018)
24. Li, M., Dong, S., Zhang, K., Gao, Z., Wu, X., Zhang, H., Yang, G., Li, S.: Deep learning intra-image and inter-images features for co-saliency detection. In: BMVC, p. 291 (2018)
25. Li, X., Yang, F., Chen, L., Cai, H.: Saliency transfer: An example-based method for salient object detection. In: IJCAI, pp. 3411–3417 (2016)
26. Li, X., Zhao, L., Wei, L., Yang, M.H., Wu, F., Zhuang, Y., Ling, H., Wang, J.: Deepsaliency: multi-task deep neural network model for salient object detection. IEEE Trans. Image Process. **25**(8), 3919–3930 (2016)
27. Li, X., Yang, F., Cheng, H., Chen, J., Guo, Y., Chen, L.: Multi-scale cascade network for salient object detection. In: Proceedings of the 25th ACM International Conference on Multimedia, ACM, pp. 439–447 (2017)
28. Li, X., Yang, F., Cheng, H., Liu, W., Shen, D.: Contour knowledge transfer for salient object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 355–370 (2018)
29. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 280–287 (2014)
30. Liu, T., Yuan, Z., Sun, J., Wang, J., Zheng, N., Tang, X., Shum, H.Y.: Learning to detect a salient object. IEEE Trans. Pattern Anal. Mach. Intell. **33**(2), 353–367 (2011)
31. Liu, Z., Shi, S., Duan, Q., Zhang, W., Zhao, P.: Salient object detection for rgb-d image by single stream recurrent convolution neural network. Neurocomputing **363**, 46–57 (2019)
32. Liu, Z., Tang, J., Zhao, P.: Salient object detection via hybrid upsampling and hybrid loss computing. In: The Visual Computer, pp. 1–11 (2019)
33. Lu, Y., Zhou, K., Wu, X., Gong, P.: A novel multi-graph framework for salient object detection. In: The Visual Computer, pp. 1–17 (2019)
34. Martin, D., Fowlkes, C., Tal, D., Malik, J.: A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Eighth IEEE International Conference on Computer Vision, 2001, ICCV 2001, Proceedings, IEEE, vol. 2, pp. 416–423 (2001)
35. Mochizuki, I., Toyoura, M., Mao, X.: Visual attention prediction for images with leading line structure. Vis. Comput. **34**(6–8), 1031–1041 (2018)
36. Qin, X., Zhang, Z., Huang, C., Gao, C., Dehghan, M., Jagersand, M.: Basnet: Boundary-aware salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7479–7489 (2019)
37. Shen, J., Peng, J., Shao, L.: Submodular trajectories for better motion segmentation in videos. IEEE Trans. Image Process. **27**(6), 2688–2700 (2018)
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556 (2014)
39. Song, H., Wang, W., Zhao, S., Shen, J., Lam, K.M.: Pyramid dilated deeper convlstm for video salient object detection. In: Proceedings of the European Conference on Computer Vision (ECCV), pp. 715–731 (2018)
40. Tang, Y., Tong, R., Min, T., Yun, Z.: Depth incorporating with color improves salient object detection. Vis. Comput. **32**(1), 111–121 (2015)
41. Wang, B., Zhang, T., Wang, X.: Salient object detection based on Laplacian similarity metrics. Vis. Comput. **34**(5), 645–658 (2018)
42. Wang, C., Zha, Z.J., Liu, D., Xie, H.: Robust deep co-saliency detection with group semantic. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 8917–8924 (2019)
43. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), pp. 136–145 (2017)
44. Wang, N., Gong, X.: Adaptive fusion for rgb-d salient object detection. arXiv preprint arXiv:1901.01369 (2019)
45. Wang, P., Wang, J., Zeng, G., Feng, J., Zha, H., Li, S.: Salient object detection for searched web images via global saliency. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, IEEE, pp. 3194–3201 (2012)
46. Wang, W., Shen, J.: Deep visual attention prediction. IEEE Trans. Image Process. **27**(5), 2368–2378 (2017)
47. Wang, W., Shen, J., Shao, L.: Video salient object detection via fully convolutional networks. IEEE Trans. Image Process. **27**(1), 38–49 (2017)
48. Wang, W., Shen, J., Yang, R., Porikli, F.: Saliency-aware video object segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **40**(1), 20–33 (2017)
49. Wang, W., Shen, J., Guo, F., Cheng, M.M., Borji, A.: Revisiting video saliency: a large-scale benchmark and a new model. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4894–4903 (2018)
50. Wang, W., Shen, J., Ling, H.: A deep network solution for attention and aesthetics aware photo cropping. IEEE Trans. Pattern Anal. Mach. Intell. **41**(7), 1531–1544 (2018)
51. Wei, L., Zhao, S., Bourahla, O.E.F., Li, X., Wu, F., Zhuang, Y.: Deep group-wise fully convolutional network for co-saliency detection with graph propagation. IEEE Trans. Image Process. **28**(10), 2052–5063 (2019)
52. Wu, Z., Su. L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3907–3916 (2019)
53. Xie, S., Tu, Z: Holistically-nested edge detection. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1395–1403 (2015)
54. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1155–1162 (2013)
55. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3166–3173 (2013)
56. Zhang, J., Ma, S., Sameki, M., Sclaroff, S., Betke, M., Lin, Z., Shen, X., Price, B., Mech, R.: Salient object subitizing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4045–4054 (2015)
57. Zhang, K., Li, T., Liu, B., Liu, Q.: Co-saliency detection via mask-guided fully convolutional networks with multi-scale label smoothing. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3095–3104 (2019)

58. Zhang, L., Dai, J., Lu, H., He, Y., Wang, G.: A bi-directional message passing model for salient object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1741–1750 (2018)
59. Zhang, Q., Lin, J., Li, W., Shi, Y., Cao, G.: Salient object detection via compactness and objectness cues. Vis. Comput. **34**(4), 473–489 (2018)
60. Zhao, J., Cao, Y., Fan, D.P., Li, X.Y., Zhang, L., Cheng, M.M.: Contrast prior and fluid pyramid integration for rgbd salient object detection. In: IEEE CVPR, pp. 1–10 (2019)
61. Zhuge, Y., Yang, G., Zhang, P., Lu, H.: Boundary-guided feature aggregation network for salient object detection. IEEE Signal Process. Lett. **25**(12), 1800–1804 (2018)



Zhengyi Liu is a Professor working with Key Laboratory of Intelligent Computing Signal Processing of Ministry of Education at Anhui University, China. Her research interests include image and video processing, computer vision and machine learning.

Qian Xiang is M.S.Candidate of Key Laboratory of Intelligent Computing Signal Processing of Ministry of Education at Anhui University, China. Her research interests include deep learning and computer vision.

Jiting Tang is M.S.Candidate of Key Laboratory of Intelligent Computing Signal Processing of Ministry of Education at Anhui University, China. Her research interests include deep learning and computer vision.

Yuan Wang is M.S.Candidate of Key Laboratory of Intelligent Computing Signal Processing of Ministry of Education at Anhui University, China. His research interests include deep learning and computer vision.

Peng Zhao is a Professor working with Key Laboratory of Intelligent Computing Signal Processing of Ministry of Education at Anhui University, China. Her research interests include information retrieval and artificial intelligence.