Salient Object Detection Using Cascaded Convolutional Neural Networks and Adversarial Learning

Youbao Tang, Xiangqian Wu, *Senior Member, IEEE*

***Abstract*—Salient object detection has received much attention and achieved great success in last several years. It is still challenging to get clear boundaries and consistent saliencies, which can be considered as the structural information of salient objects. A popular solution is to conduct some post-processes (e.g. CRF) to refine these structural information. In this paper, a novel cascaded convolutional neural networks based method is proposed to implicitly learn these structural information via adversarial learning for salient object detection (denoted CCAL). A cascaded CNNs model is first designed as a generator G, which consists of an encoder-decoder network for global saliency estimation and a deep residual network for local saliency refinement. It is hard to explicitly learn such structural information due to the limitation of frequently-used pixel-wise loss functions. Instead, a discriminator D is then designed to distinguish the real salient maps (i.e. ground truths) from the fake ones produced by G, based on which an adversarial loss is introduced to optimize G. G and D are trained in a fully end-to-end fashion by following the strategy of Conditional Generative Adversarial Networks (CGAN) to make G well learn the structural information. At last, G is able to produce high quality salient maps without requiring any post-process to fool D. Experimental results on eight benchmark datasets demonstrate the effectiveness and efficiency (about 17 fps on GPU) of the proposed method for salient object detection.**

***Index Terms*—Salient object detection, cascaded convolutional neural networks, conditional generative adversarial networks, adversarial learning.**

I. Introduction

As an important and challenging topic in computer vision, salient object detection aims to locate the objects which attract our attention in natural images by giving them large salient values. Recently, there has been a lot of interest in developing different kinds of models for salient object detection, due to their advantages in a variety of computer vision applications such as image segmentation [1], visual tracking [2], and object recognition [3], etc.

In the past two decades, a large number of salient object detection approaches [4]–[45] have been developed. Most of them have been investigated by Borji et al [29]. The early salient object detection approaches [4]–[16], [34] focus on extracting discriminative local and global hand-crafted features from pixels or regions to represent their properties. Then some machine learning algorithms are used to compute the salient scores according to the extracted features for salient object detection. Although these approaches have gotten great achievements, there still exist many problems which need to be solved. For example, the hand-crafted features are difficult to capture the semantic and structural information of salient objects in images, which is important for salient object detection. And to further extract more powerful features manually is a tough mission for performance improvement of salient object detection.

Inspired by the successes of convolutional neural networks (CNNs) in a wide range of computer vision tasks (e.g. image classification [30], object detection [31], and semantic image segmentation [46]) and many other fields [47]–[53], many researchers [19]–[21], [23]–[26], [32], [33], [36]–[39], [44], [45] make their efforts to use CNNs for salient object detection and achieve state-of-the-art performance, since CNNs have strong ability to automatically learn high-level feature representations of salient object, so as to successfully avoid the problems of hand-crafted features. Here, we briefly discuss some CNN based salient object detection approaches. Zhao *et al*. [21] propose a multi-context deep learning framework to consider the global and local context of salient objects. Wang *et al*. [19] integrate both local estimation and global search for salient object detection by training two deep neural networks. Li *et al*. [20] employ CNN to extract multiscale deep features for saliency computation. And then they propose a deep contrast network to combine a pixel-level stream and a segment-wise stream for saliency estimation [26]. Wang *et al*. [23] develop a recurrent fully convolutional network to incorporate saliency prior knowledge for more accurate inference. Tang *et al*. [24] use a CNN to fuse pixel-level and region-level saliencies estimated by two CNNs to get the final saliency prediction. Liu *et al*. [25] design a deep hierarchical network to learn a coarse global prediction and then refine the salient map hierarchically and progressively. Zhang *et al*. [37] design a framework to aggregate multi-scale and multi-level convolutional features for salient object detection. And they also develop a network to learn the deep uncertain convolutional features to enhance the robustness and accuracy of saliency detection [38]. Wang *et al*. [39] provide a stage-wise saliency refinement framework to gradually get accurate saliency detection results. Hou *et al*. [33] introduce short connections to the skip-layers within HED architecture to get rich multi-scale features for salient object detection. Luo *et al*. [36] add a boundary loss team to training loss function to penalize errors on the boundary for saliency detection improvement. Wang *et al*. [32] develop a weakly supervised learning method for saliency detection

Youbao Tang and Xiangqian Wu (corresponding author) are with the School of Computer Science and Technology, Harbin Institute of Technology (HIT), Harbin, 150001 China (e-mail: tybxiaobao@gmail.com, xqwu@hit.edu.cn).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2900908, IEEE Transactions on Multimedia

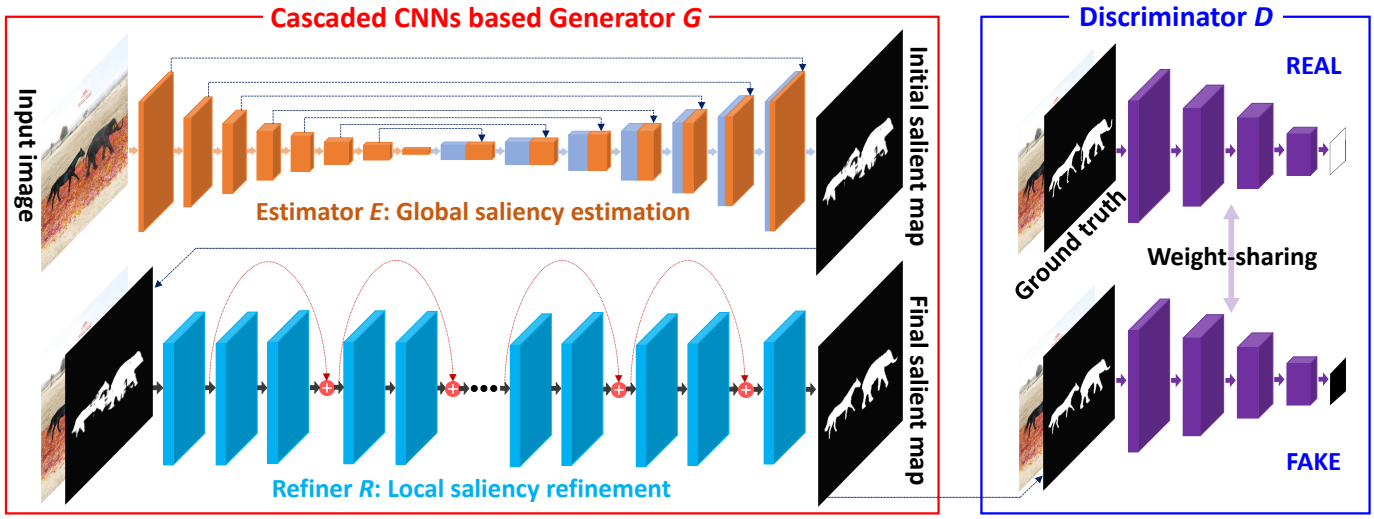


Fig. 1. Overview of the proposed cascaded convolutional networks based method for salient object detection.

using image-level tags only. Chen *et al*. [35] propose two-stream CNNs for fixation prediction and semantic perception, respectively. These two streams are fused and jointly fine-fined for salient object detection. Zhang *et al*. [45] propose a bi-directional message passing model to integrate multilevel features extracted adopting a multi-scale context-aware feature extraction module for salient object detection. Wang *et al*. [44] detect the salient objects by designing a recurrent localization network and a boundary refinement network for global detection and local refinement respectively. Despite these approaches employ CNNs to get remarkable successes in salient object detection, there still exist some problems. For example, the strategies of multiple-stages reduce the efficiency. And the explicit pixel-wise loss functions used by these methods for training cannot well reflect the structural information of salient objects. Hence, there is still a large space for improvement.

In this work, we propose a novel salient object detection method (named CCAL) based on cascaded convolutional neural networks and adversarial learning. For the cascaded convolutional neural networks design, we develop an encoder-decoder network for global saliency estimation and a deep residual network for local saliency refinement with a cascaded manner, which can be considered as a generator. With this generator, the salient object detection performance is improved gradually. Different from previous CNN based approaches [19]–[21], [23]–[26], [32], [33], [36]–[39], an adversarial loss is introduced by design a discriminator in this work, which attempts to explore a structured loss function. And following the Conditional Generative Adversarial Networks (CGAN) [54], the adversarial learning strategy is employed to help the generator to well learn the structural information (i.e. clear boundary and consistent saliencies) of salient objects. With the above considerations, the proposed method CCAL can produce good salient object detection results with clear boundaries and consistent saliencies. With the trained model, the input image is fed to the generator to get the final salient map with a single feedforward pass, which guarantees the efficiency of the proposed method.

Compared with our CCAL, RFCN [23], DGRL [44] and SalGAN [55] are three most related works. RFCN and DGRL also detect the salient objects with a coarse-to-fine manner. RFCN needs an extra input channel (i.e. a saliency prior map) for the coarse saliency estimation and stacks the encoder-decoder networks (EDN) for saliency refinement. DGRL builds a ResNet-50 based recurrent localization network to localize the salient object globally and uses a boundary refinement network containing 9 convolutional layers for saliency refinement. Both RFCN and DGRL optimize the models with pixel-wise loss functions. Our CCAL only uses the input of RGB image for coarse global saliency estimation, designs a deep residual network (DRN) for local saliency refinement and employs the adversarial learning strategy for training. Although SalGAN [55] has used CGAN for fixation prediction, it uses an EDN as generator while our CCAL proposes a cascaded CNNs model containing both an EDN and a DRN as generator and successfully uses CGAN for salient object detection.

In summary, this paper makes the following main contributions:

(1) We design a novel network framework for salient object detection which contains two convolutional neural networks combined with a cascaded manner. They focus on global saliency estimation and local saliency refinement, respectively. With the gradual help, the detection results are improved progressively.
(2) We employ CGAN for salient object detection, where the adversarial loss is introduced to implicitly learn the structural information (i.e. clear boundaries and consistent saliencies) for performance improvement.
(3) We evaluate the proposed method over eight benchmark datasets. Comprehensive experimental results demonstrate that the proposed method is able to produce high quality salient maps with clear boundaries and consistent saliencies, and significantly outperforms state-of-the-art approaches.

II. The Architecture of the Proposed Network

There are two components in the proposed salient object detection model, i.e. the generator G and the discriminator D, as shown in Fig. 1. The generator is a cascaded convolutional neural networks based model, where the input is a RGB image and the output is the predicted salient map for each input image. To achieve the best performance, we develop an encoder-decoder network (called global saliency estimator E) for global saliency estimation and a deep residual network (called local saliency refiner R) for local saliency refinement. The generator can be simply updated with pixel-wise loss functions. The predicted salient maps may have some blurred boundaries and inconsistent saliencies. It is necessary to utilize the powerful ability of discriminator to improve the predictions. Therefore, a real and a fake discriminator are designed and an adversarial loss is introduced to help the generator to well learn the structural information of salient objects. In this section, the details of these network architectures will be described, and the reasons of their designs also will be explained.

A. Cascaded CNNs based Generator $\boldsymbol{G}$

1) Global Saliency Estimator $\boldsymbol{E}$: Salient object detection can be treated as a pixel labeling problem by assigning large value (e.g. 1) for salient objects and small one (e.g. 0) for non-salient regions. For the pixel labeling tasks (e.g. semantic image segmentation [46], [56], also including salient object detection [19]–[21], [23]–[26], [32], [33], [36]–[39], [57]), the state-of-the-art performances are hold by CNN based approaches. Most of them are modified from VGGNet [30] or encoder-decoder networks [56].

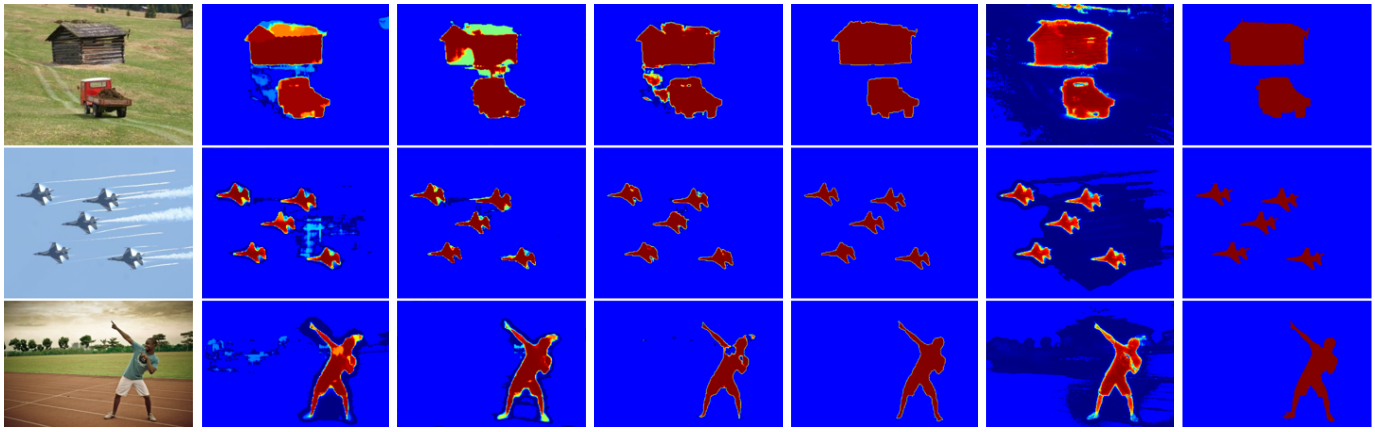
Inspired by the successes of encoder-decoder networks, this work also builds an encoder-decoder network (named global saliency estimator E) for initial salient map estimation, which contains two parts, i.e. encoder and decoder. The duty of encoder is to reduce the size of feature maps through feedforwarding to learn global features of salient objects. Instead of utilizing pooling operations, this work uses convolutions with a large stride to downsample the feature maps. Specifically, we use convolution with kernel size 4×4 and stride 2 to replace the combination of convolution with kernel size 3×3 and stride 1 and pooling with size 2×2 and stride 2, which is a classical setting in VGGNet [30]. And to make the final extracted features of encoder capture the global information of salient objects, multiple convolutional layers are stacked. Here, there are $n_1 = 8$ convolutional layers in our encoder, and the number of convolution kernels in each layer are 64, 128, 256, 512, 512, 512, 512, 512, respectively.

For decoder, it performs an opposite process of encoder by enlarging the size of feature maps. As done by [56], [58], [59], the deconvolutional operation with kernel size 4×4 and stride 2 is adopted to upscale the feature maps in this work. Besides, we also use skip connections to combine the high-level features from decoder and low-level features from encoder to boost feature learning as done by [58], [59]. To be specific, the skip connections are added between each convolutional layer i and its corresponding symmetric deconvolutional layer $n_2 - i$ by simply concatenating their outputs in channel direction, where $n_2 = n_1 = 8$ is the number of deconvolutional layers. The number of kernels in each layer are 512, 512, 512, 512, 256, 128, 64, 1, respectively. All layers are followed by a ParametricReLU activation function [60] and a batch normalization layer [61]. And the last layer is followed by a tanh activation function. The top of red rectangle in Fig. 1 shows the architecture of the encoder-decoder network (i.e. saliency estimator E), which is similar with UNet [58]. From Fig. 1, given an input image, the output of E is a probability map which has the same size with the input image and is considered as the initial salient map, in which the salient objects have been highlighted and the backgrounds have been suppressed.

2) Local Saliency Refiner $\boldsymbol{R}$: In the initial salient maps, although the salient objects have been highlighted and the backgrounds have been suppressed, there still exist some local regions where the saliencies are poorly estimated (as shown in Fig. 1). It is necessary to leverage the information provided by initial salient map to correct these poor estimations. Therefore, we design a deep residual network (named local saliency refiner R) for local saliency refinement, where the inputs are the combination of the RGB image and the initial salient map produced by saliency estimator E and the output is the refined salient map as the final result for performance evaluation.

To precisely extract features from local regions, it is not a good design to downsample the features gradually as done by the encoder of saliency estimator E, but the receptive field of the designed network should be large enough to cover the local regions. Therefore, in this work, we develop a deep fully convolutional neural network to achieve this goal, which contains 34 convolutional layers. Directly stacking these layers tends to suffer from optimization difficulty caused by vanishing gradients. Instead, we adopt residual block design with identical layout proposed by Gross and Wilber [62] to alleviate this problem. Specifically, 16 residual blocks are used and each residual block consists of two convolutional layers with 64 3×3 kernels followed by batch-normalization layers [61] and ParametricReLU [60] activation function. And two another convolutional layers with 64 3×3 kernels are also employed, one is used at the beginning followed by ParametricReL and the other one is used at the end followed by a convolutional layer with one 1×1 kernel and a tanh activation function. The stride and pad of all convolutional operations are 1. Therefore, all extracted features have the same size with the input and the largest receptive field is 67, which makes the deep residual network contain abundant contextual information for local saliency refinement. The bottom of red rectangle in Fig. 1 shows the architecture of the deep residual network (i.e. saliency refiner R).

With the power of the proposed saliency refiner R, the poor estimations of E can be corrected. From Fig. 1, compared with the initial salient map estimated by E, the refined result (the final salient map) is closer to the ground truth, because some local poor or wrong saliencies on both salient objects and backgrounds have been corrected by R. With investigation, we found that the result would be a little worse when R has 9 residual blocks.



(a) (b) (c) (d) (e) (f) (g)

Fig. 2. Three examples of salient object detection results by the proposed method. (a) Input images. (b) Detection results produced by only using E. (c) Detection results produced by the generator G including E and R without the discriminator D for training (denoted $E+R$). (d) Detection results produced by global saliency estimator E with the discriminator D for training (denoted $E+D$). (e) Detection results produced by the generator G including E and R with the discriminator D for training (denoted $E+R+D$). (f) Detection results produced by E and refined by CRF [63] (denoted $E+CRF$). (g) Ground truths. The original salient maps and ground truths are gray images. To better visualize the differences between detected salient maps and ground truths, we colorize them with pseudo-colors, which also is applied in other figures. Better viewed in color.

B. Discriminator D

As above descriptions, given an input image I, its final salient map X generation process can be formulated as $X = G(I) = R(I, E(I))$. The previous CNN based salient object detection approaches [19]–[21], [23]–[26], [32], [33], [36]–[39] use some explicit pixel-wise loss functions (e.g. ℓ_1 and ℓ_2 norm) to calculate the pixel-wise errors between X and its corresponding ground truth Y. These loss functions operate on pixel level and can well measure the pixel-wise differences, but they are difficult to capture high level contextual and structural details, resulting in producing blurred results [64]. So introducing a structured loss function can handle this situation. The discriminator in Generative Adversarial Network (GAN) [65] can be considered as an attempt to explore a structured loss function. Therefore, to make the generator G well learn the structural information of salient objects, we design a discriminator D, whose role is to distinguish the fake salient maps produced by the generator G from the real ones (i.e. ground truths) by following the strategy of conditional GAN (CGAN) [54]. CGAN is a conditional version of GAN.

The blue rectangle in Fig. 1 gives the architecture of D, which is a Siamese network. When the inputs are I and X (or Y), D try to discriminate X (or Y) as a generated salient map (or ground truth). In previous approaches [64], [65], D usually outputs a single value 0 or 1 to make the judgment for the entire image. In this work, to make the model capture the local structures of salient objects, D do the judgment for every $N \times N$ local image patch as done by [66]. Here, we set $N = 31$. Hence, D consists of five convolutional layers, as shown in Fig. 1. The first four layers have the same network configurations with the first four layers of the encoder in saliency estimator E. The last convolutional layer with a 1×1 kernel is followed by sigmoid activation function. Based on the discriminator, an adversarial loss is introduced to make the generator further learn the structural information of salient objects, thus more confident results are produced. Here, please notice that we expect that the discriminator D can discriminate all local image patches as ones or zeros. When all patches can be correctly discriminated, the entire image should also be correct. When using a single scalar value 1 or 0 to make a global judgment, it cannot guarantee all local patches are discriminated properly, since the global judgment is given by linearly or nonlinearly fusing all local judgments. Therefore, the global judgment is different with multiple local judgments which can make the local patches of a salient map clearer and sharper.

As above descriptions, the whole architecture of the proposed network has been introduced, which is a fully convolutional neural network. Fig. 2 gives three examples of salient object detection results produced by different model configurations to intuitively verify the benefits of our local saliency refiner R and discriminator D. Given the input images (as shown in Fig. 2(a)), to verify the effectiveness of refiner R and discriminator D designs, we train the generator G which includes R or not with or without using discriminator D. Their corresponding saliency detection results are given in Fig. 2(c)-(e), respectively. From Fig. 2(b) and Fig. 2(c) (or Fig. 2(d) and Fig. 2(e)), we can see that some local poor or wrong saliencies on both salient objects and backgrounds have been corrected by the refiner R. From Fig. 2(b) and Fig. 2(d) (or Fig. 2(c) and Fig. 2(e)), we can see that the results with discriminator D (as shown in Fig. 2(d) and Fig. 2(e)) get clearer boundaries and more consistent saliencies than the ones without discriminator D (as shown in Fig. 2(b) and Fig. 2(c)). As we know that CRF [63] is a popular model to refine structural information. Here, we also use the CRF model as a post-process to refine the results produced by E (see Fig. 2(f)). From Fig. 2(e) and Fig. 2(f), we can see that the boundaries are clearer and the saliencies are more consistent in Fig. (e) than the ones in Fig. (f), suggesting that our refinement module consisting of R and D is more powerful than CRF for saliency refinement. These

results visually demonstrate the effectiveness of our refiner R and discriminator D for salient object detection.

III. Model Optimization

After designing the architecture, next step is to train the proposed model CCAL for salient object detection. First of all, the object function needs to be defined before model training, which is critical for the performance of our CCAL model. In this work, the objective function is defined as follows:

$$\min_D L(D) \tag{1}$$

$$\min_G L(G) \tag{2}$$

where $L(D)$ is the loss function for discriminator D optimization, and $L(G)$ is the loss function for generator G optimization. Compared with the sigmoid cross entropy loss function, the least squares loss function [67] which makes the model more stable is used as $L(D)$ in this work defined by

$$L(D) = \frac{1}{2}\mathbb{E}[(D(G(I), I)^2] + \frac{1}{2}\mathbb{E}[(D(Y, I) - \mathbf{1})^2] \tag{3}$$

The loss function $L(G)$ is formulated as the weighted sum of a content loss L_C and an adversarial loss L_A as follows:

$$L(G) = \lambda \cdot L_C + (1 - \lambda) \cdot L_A \tag{4}$$

where $\lambda = 0.99$ is a weight to control the importance of content loss L_C and adversarial loss L_A. For the content loss L_C, it is used to compute the pixel-wise error between the ground truth and the salient maps generated by estimator E and refiner R. So the content loss L_C is the weighted sum of an estimator loss L_E and a refiner loss L_R as follows:

$$L_C = \gamma \cdot L_E + (1 - \gamma) \cdot L_R \tag{5}$$

where γ is a weight to control the importance of L_E and L_R. In this work, we set $\gamma = 0.5$ to equally consider the contribution of L_E and L_R. For most of images, the number of salient object pixels and non-salient pixels are heavy imbalance. As done by other works, the cross-entropy loss function defined in [68] is used to balance the loss between salient and non-salient classes as follows:

$$\begin{aligned} L_E = &- \alpha \sum_{i=1}^{|Y_+|} \log P(y_i = 1 | E(I)) \\ &- (1 - \alpha) \sum_{i=1}^{|Y_-|} \log P(y_i = 0 | E(I)) \end{aligned} \tag{6}$$

where $\alpha = \frac{|Y_-|}{|Y_+|+|Y_-|}$, $|Y_+|$ and $|Y_-|$mean the number of salient pixels and non-salient pixels in ground truth. We can use $R(I, E(I))$ to replace $E(I)$ in Equation 6 to compute the loss L_R.

Since G tries to generate high quality salient map to fool D, the adversarial loss L_A is defined as

$$L_A = \frac{1}{2}\mathbb{E}[(D(G(I), I) - \mathbf{1})^2] \tag{7}$$

From $L(G)$, we can see that the task of G is not only to generate salient map which is close to the ground truth in pixel-level but also to fool D.

The **MSRA10K** dataset [11] is used to train our CCAL model, which contains 10,000 images. To increase the diversity of training samples, for each image, we randomly generate a number of synthetic images by using the following data augmentation technique: (1) Select a salient object or not. (2) If selected, randomly rotate and resize it, and then randomly put it into the background region. We can repeat Step 2 to generate multiple salient objects with variant scales. (3) Randomly crop an image with size 256×256 from above constructed image and randomly flip it to form a training sample. The above processes are also simultaneously performed on the ground truths to produce the synthetic ground truths corresponding to their synthetic images.

MSRA10K dataset is split into a training set containing 9,000 images and a validate set containing 1,000 images. We use the training set and the above data augmentation technique to generate samples for model training, and set the batchsize as 5. To optimize our defined objective function, we alternate training of G and D, one step on G, and three steps on D. The framework of the proposed method is implemented based on the publicly available Tensorflow library [69]. The Adam optimizer [70] is used and the learning rate is 0.002. It takes about 32h to train the model from scratch until convergence after $\sim$15 epochs using a Titan X Pascal GPU. The number of augmented training samples is $9000 \times 15 = 135000$. After training, the model which gets the best performance on the validate set is selected and used for performance evaluation on all test datasets.

IV. Experiments

A. Datasets and Evaluation Criteria

The performance evaluation is conducted on eight standard benchmark datasets: SED1 [71], SED2 [71], ECSSD [4], PASCAL-S [72], HKU-IS [20], SOD [73], DUT-OMRON [74], and DUTS-TE [32].

SED1 and **SED2** [71] contain 100 images with one and two salient object, respectively, in which objects have largely different sizes and locations.

ECSSD [4] contains 1,000 images with complex backgrounds.

PASCAL-S [72] contains 850 images with multiple complex objects and cluttered backgrounds. Different salient objects are labeled with different saliencies. This dataset is arguably one of the most challenging saliency data sets without various design biases (e.g., center bias and color contrast bias).

HKU-IS [20] contains 4447 challenging images with multiple disconnected salient objects, overlapping the image boundary or low color contrast.

SOD [73] which is selected from the BSDS dataset [75] contains 300 images. Many images have multiple objects either with low contrast or touching the image boundary.

DUT-OMRON [74] has 5,168 high quality images. Images of this dataset have one or more salient objects and relatively complex background. Thus this dataset is more difficult and challenging, and provides more space of improvement for related research in saliency detection.

DUTS-TE [32] has 5,019 images with high quality pixel-wise annotations, which is the testing set of **DUTS** [32]. **DUTS**

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2900908, IEEE Transactions on Multimedia

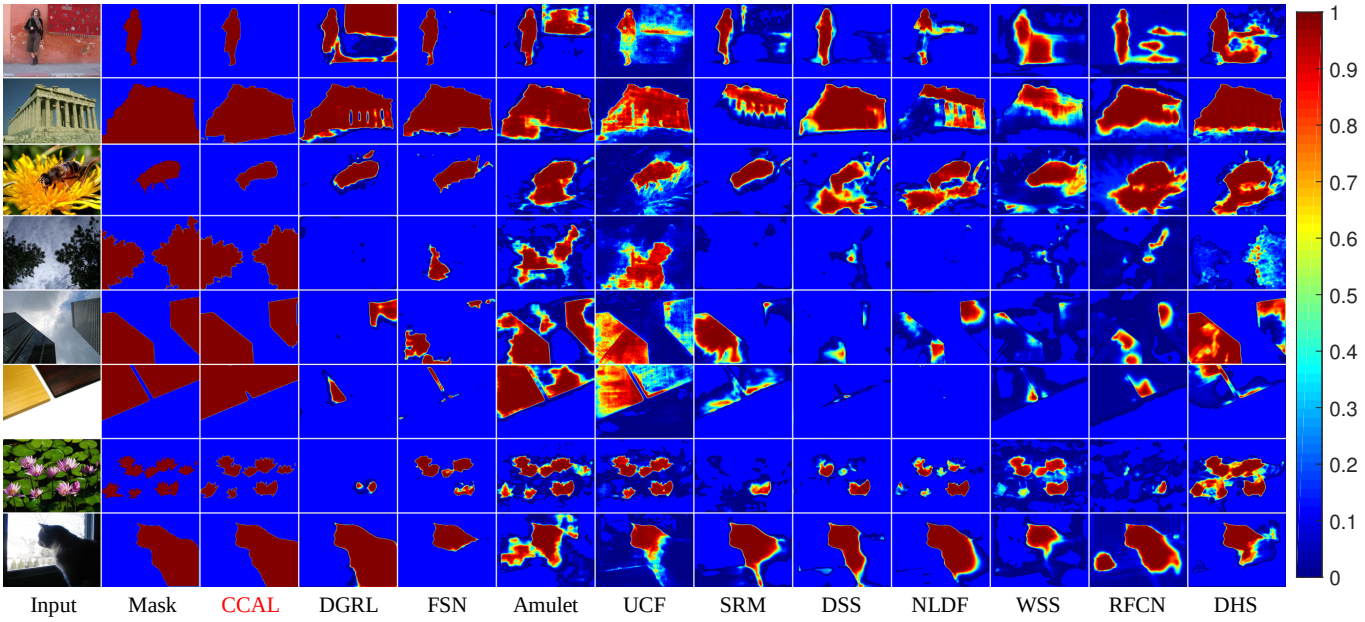


Fig. 3. Visual Comparisons of different saliency detection approaches vs. our method (CCAL) in various challenging scenarios.

also includes another 10,553 images for training. It contains very challenging scenarios for saliency detection. All datasets provide the corresponding ground truths in the form of accurate pixel-wise human-marked labels for salient objects.

Six popular criteria are used for performance evaluation, i.e. precision and recall curve (denoted PR curve), F-measure (denoted F_β) and threshold curve (denoted FT curve), mean F_β, weighted F-measure (denoted wF_β) [76], structural similarity measure (SSM) [77], and mean absolute error (MAE).

For a given salient map S, its binary mask B can be obtained using a threshold. Then the precision and recall can be calculated according to its corresponding ground truth GT by $\frac{|B \cap GT|}{|B|}$ and $\frac{|B \cap GT|}{|GT|}$. The PR curve is drawn using the thresholds in the range [0, 255]. F_β is the overall performance measurement computed by the weighted harmonic of precision and recall:

$$\mathrm{F}_\beta = \frac{(1+\beta^2) \times Precision \times Recall}{\beta^2 \times Precision + Recall} \tag{8}$$

where we set $\beta^2 = 0.3$ as used by other approaches.

The weighted F-measure wF_β [76] which has offered an intuitive generalization of the F_β is also used for evaluation in this work.

The structural similarity measure (SSM) proposed recently in [77] is able to simultaneously evaluate region-aware and object-aware structural similarity between S and GT.

The mean absolute error (MAE) is computed by

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

where W and H are the width and height of the image, S and GT are normalized to [0, 1].

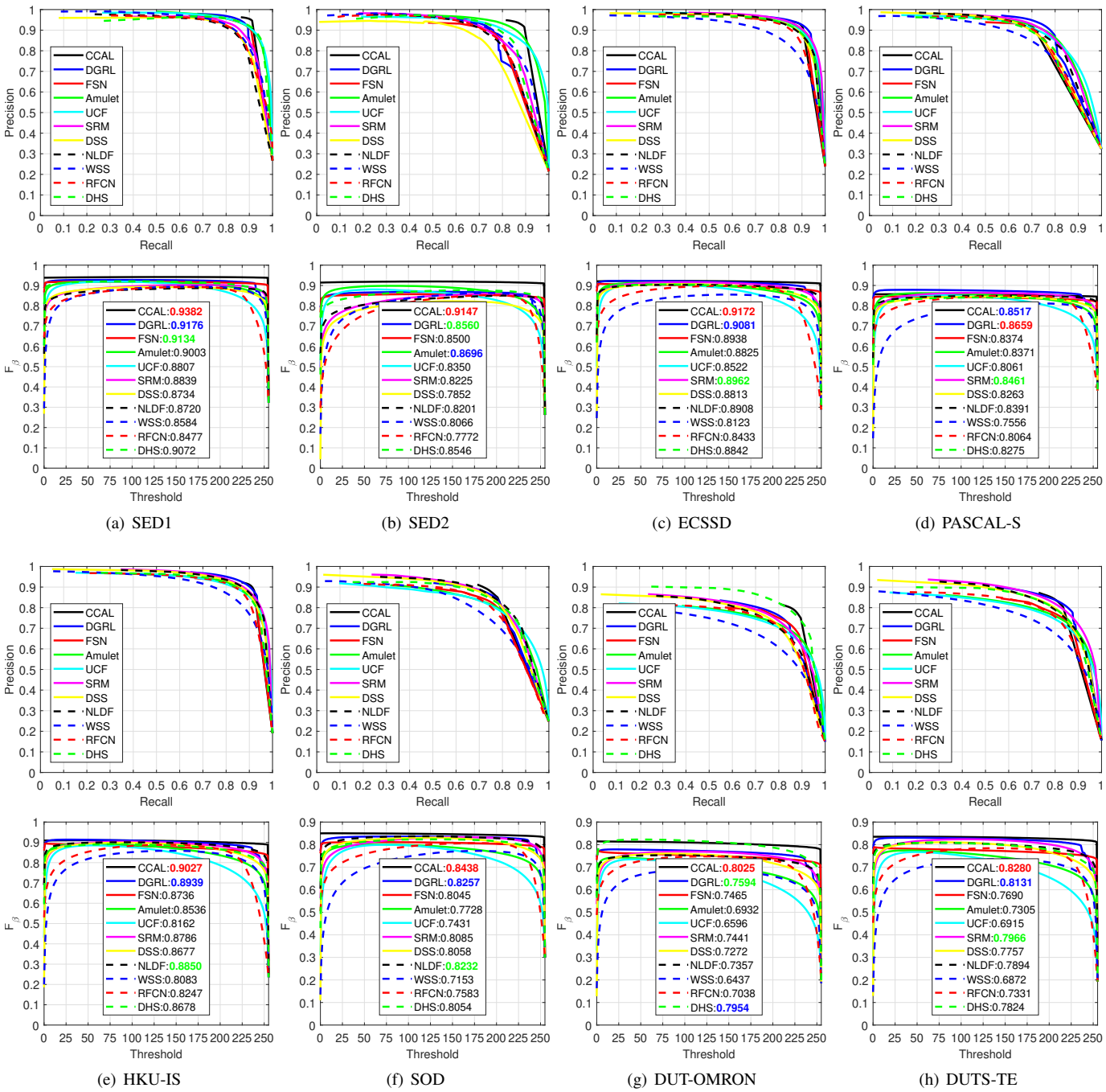
B. Comparison with State-of-the-arts

The performance of the proposed method (CCAL) is compared with ten state-of-the-art CNN based salient object detection approaches on eight test datasets, including DGRL [44], FSN [35], Amulet [37], UCF [38], SRM [39], DSS [33], NLDF [36], WSS [32], RFCN [23], and DHS [25]. For fair comparison, the source codes of these approaches released by the authors are used for test with recommended parameter settings in this work.

Figure 3 provides a visual comparison of our method CCAL with respect to other approaches. From Fig. 3, our method gets the best detection results which are much close to the ground truth in various challenging scenarios. To be specific, (1) the proposed method not only highlights the right salient object regions clearly, but also well suppresses the saliencies of background regions, so as to produce the detection results with higher contrast between salient objects and background than other approaches. (2) With the help of the adversarial loss, the proposed method is able to generate the salient maps with clear boundaries and consistent saliencies, compared with the blurred results obtained by other approaches which only use the explicit pixel-wise loss functions.

Figure 4 provides the quantitative evaluation results of the proposed method and other approaches on all test datasets in terms of PR curve, FT curve and mean F_β criteria. From Fig. 4, we can conclude that the proposed method almost consistently assigns the largest saliency to the detected salient regions and the smallest saliency to backgrounds, which can be proved by (1) as the threshold increasing during criteria computation, our smallest recalls on all datasets are much larger than others (seeing the top figure of each dataset in Fig. 4), (2) our F-measures change very little and are almost the best compared with others (seeing the bottom figure of each dataset in Fig. 4), so the proposed method gets the best mean F-measure (seeing the numbers in the bottom figure of each dataset in Fig. 4).

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2900908, IEEE Transactions on Multimedia



(a) SED1 (b) SED2 (c) ECSSD (d) PASCAL-S

(e) HKU-IS (f) SOD (g) DUT-OMRON (h) DUTS-TE

Fig. 4. Quantitative results of all approaches on eight test datasets in terms of PR curves, FT curves, and mean F_β. For each dataset, the top figure shows the PR curves while the bottom figure shows the FT curves and the mean F_β values in the legends.

Table I lists the results of all approaches in terms of wF_β, SSM and MAE over all test datasets. From Table I, we can see that the proposed method almost achieves the best scores of wF_β, SSM and MAE on all test datasets except the wF_β and SSM on PASCAL-S, DUT-OMRON and DUTS-TE datasets. After investigation, we find the possible reason of worse performance on PASCAL-S is that when the images in PASCAL-S contain multiple salient objects, the salient objects are labeled with different salient values in ground truths, but our method still assigns coherent saliency to them (seeing the second row of Fig. 5). DHS [25] gets comparable and better performance than our method on DUT-OMRON. That's because DHS uses a large part of images in DUT-OMRON dataset for training and its performance is evaluated on the rest 1,200 images as done in [25] here. DGRL [44] is trained on the training set of DUTS, thus it gets better wF_β than our method in DUTS-TE. In addition, Table I also lists the model sizes of different methods. Compared with the second best approach (*i.e.* DGRL), our method utilizes a much smaller model to get the best overall performance, suggesting that the performance improvement of the proposed method is from the new designed components rather than larger capacity of the

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2900908, IEEE Transactions on Multimedia

TABLE I
THE MODEL SIZES OF DIFFERENT SALIENT OBJECT DETECTION APPROACHES AND THEIR wF_β, SSM AND MAE ON ALL TEST DATASETS (**RED**, **BLUE**, AND **GREEN** TEXTS RESPECTIVELY INDICATE RANK 1, 2, AND 3).

Method	Size	SED1			SED2			ECSSD			PASCAL-S		
		wF_β	SSM	MAE	wF_β	SSM	MAE	wF_β	SSM	MAE	wF_β	SSM	MAE
CCAL	43.0M	**0.9140**	**0.9070**	**0.0372**	**0.8930**	**0.8861**	**0.0370**	**0.8868**	**0.8952**	**0.0415**	0.7403	0.7568	**0.0709**
DGRL	161.8M	0.8673	0.8860	**0.0548**	0.7658	0.7919	**0.0752**	**0.8791**	**0.8939**	**0.0447**	**0.7801**	**0.7959**	**0.0742**
FSN	47.5M	**0.8765**	0.8839	**0.0525**	0.7843	0.8034	0.0859	**0.8595**	0.8798	**0.0554**	0.7392	0.7564	0.0924
Amulet	**33.1M**	0.8564	0.8894	0.0618	**0.8253**	**0.8448**	**0.0644**	0.8396	0.8906	0.0607	**0.7547**	**0.7937**	0.0997
UCF	**29.5M**	0.8320	**0.8913**	0.0737	**0.7863**	**0.8395**	0.0767	0.7879	0.8797	0.0797	0.7129	**0.7878**	0.1268
SRM	53.2M	0.8099	0.8496	0.0770	0.6984	0.7558	0.0933	0.8495	**0.8910**	0.0564	**0.7445**	0.7815	**0.0835**
DSS	62.2M	0.8003	0.8475	0.0813	0.6666	0.7435	0.1055	0.8318	0.8792	0.0646	0.7108	0.7515	0.1016
NLDF	106.5M	0.7815	0.8216	0.0913	0.6837	0.7447	0.1053	0.8354	0.8698	0.0658	0.7267	0.7559	0.0979
WSS	**14.7M**	0.7656	0.8296	0.1005	0.6916	0.7651	0.1007	0.7113	0.8081	0.1059	0.6182	0.7043	0.1395
RFCN	134.7M	0.7438	0.8371	0.1038	0.6297	0.7418	0.1137	0.7253	0.8561	0.0972	0.6573	0.7576	0.1176
DHS	94.0M	**0.8683**	**0.8921**	0.0550	0.7599	0.7916	0.0810	0.8368	0.8798	0.0621	0.7123	0.7522	0.0918
Method	**Size**	**HKU-IS**			**SOD**			**DUT-OMRON**			**DUTS-TE**		
		wF_β	SSM	MAE	wF_β	SSM	MAE	wF_β	SSM	MAE	wF_β	SSM	MAE
CCAL	43.0M	**0.8794**	**0.8974**	**0.0329**	**0.7412**	**0.8026**	**0.0827**	**0.7243**	**0.8259**	**0.0418**	**0.7315**	**0.8367**	**0.0474**
DGRL	161.8M	**0.8649**	**0.8968**	**0.0374**	**0.7409**	**0.7962**	**0.0887**	**0.6968**	**0.8097**	**0.0632**	**0.7533**	**0.8356**	**0.0512**
FSN	47.5M	**0.8454**	0.8763	**0.0442**	0.7239	0.7771	0.1070	0.6961	0.8030	0.0652	0.7106	0.8011	0.0656
Amulet	**33.1M**	0.8100	0.8814	0.0531	0.6842	0.7763	0.1248	0.6256	0.7805	0.0976	0.6534	0.7956	0.0851
UCF	**29.5M**	0.7476	0.8646	0.0749	0.6370	0.7659	0.1527	0.5646	0.7580	0.1316	0.5875	0.7727	0.1170
SRM	53.2M	0.8310	**0.8838**	0.0469	0.6880	0.7690	0.1072	0.6577	0.7977	0.0694	**0.7147**	**0.8244**	**0.0587**
DSS	62.2M	0.8194	0.8786	0.0509	0.7094	**0.7865**	0.1066	0.6430	0.7892	0.0745	0.6975	0.8168	0.0651
NLDF	106.5M	0.8353	0.8762	0.0490	**0.7264**	0.7837	**0.1032**	0.6337	0.7703	0.0799	0.7028	0.8051	0.0654
WSS	**14.7M**	0.7136	0.8219	0.0796	0.5750	0.7031	0.1471	0.5253	0.7302	0.1100	0.5584	0.7370	0.0999
RFCN	134.7M	0.7051	0.8544	0.0804	0.6139	0.7503	0.1368	0.5751	0.7763	0.0913	0.5949	0.7851	0.0870
DHS	94.0M	0.8158	0.8703	0.0529	0.7007	0.7758	0.1077	**0.7333**	**0.8444**	**0.0491**	0.7015	0.8109	0.0657

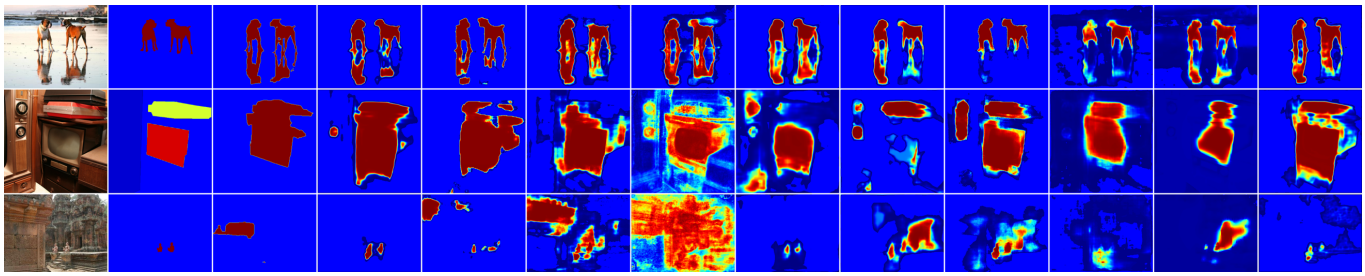


Fig. 5. Failure salient object detection examples of different approaches.

model.

Besides, from Fig. 4 and Table I, we also observe that our method gets larger improvement compared with the best existing approach on SED2 dataset where all images contain two salient objects than others. The possible reason is that the proposed data augmentation technique can generate a number of images containing multiple salient objects, based on which the proposed CCAL model can be well trained to tackle multiple salient object detection. The results presented in Fig. 4 and Table I demonstrate the effectiveness of the proposed method for salient object detection.

We also test the average runtime of the proposed method over ECSSD dataset. The test is taken on a PC with an Intel i7-5930K CPU, a TITAN X Pascal GPU, and 64GB memory. The average runtime is about 0.06 second.

Figure 5 shows some failure examples of the proposed method. For example, when the salient objects have clear reflections (seeing the first row of Fig. 5), the proposed method detect the reflections as salient objects too, but the human visual system can easily focus attention on the real objects. When the salient objects have different saliencies (seeing the second row of Fig. 5), the proposed method cannot assign different values to them and distinguish them. When the salient objects are very small and have similar colors with backgrounds (seeing the third row of Fig. 5), it is difficult to correctly detect them with our method. The other approaches almost also fail to do the salient object detection correctly in these cases.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2019.2900908, IEEE Transactions on Multimedia

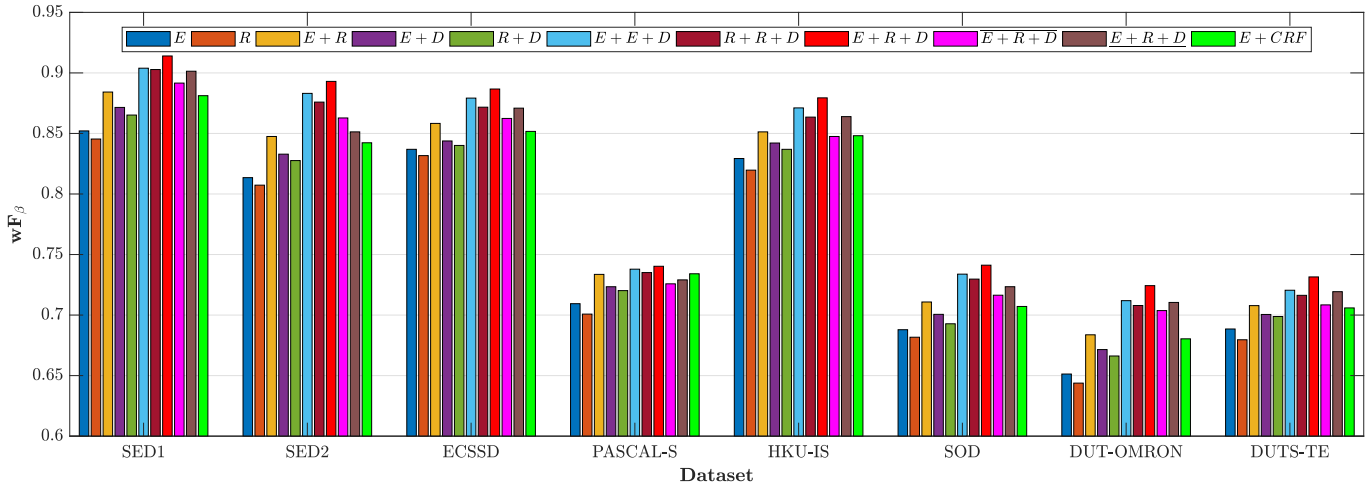


Fig. 6. Salient object detection results on all test datasets in terms of wF_β when using different component combinations.

C. Ablation Study

As described in Section 2, the proposed method contains two main parts, i.e. the cascaded CNNs based generator G (including the initial saliency estimator E and the saliency refiner R) and the discriminator D. To investigate the importance of our cascaded CNNs design and the adversarial learning, we conduct the following experimental comparisons: E, R, $E+R$, $E+D$, $R+D$, $E+E+D$, $R+R+D$, $\overline{E+R+D}$, $\underline{E+R+D}$, $E+CRF$ and $E+R+D$. For instance, E or R means only E or R is used and trained with only loss L_E or L_R, $E+R$ means E and R are cascaded and trained with loss L_E and L_R, $E+E+D$ means the global saliency estimator E is also used for saliency refinement and then combined with D for training with loss L_E, L_E and L_A, $\overline{E+R+D}$ means all components are used with the same way as the proposed method $E+R+D$ except using only the adversrial loss L_A for training, $\underline{E+R+D}$ means the proposed method $E+R+D$ is trained without using our data augmentation technique, and $E+CRF$ means using a CRF model [63] as a post-process to refine the initial salient maps produced by E. The CRF's parameters are determined by conducting cross validation on the validation set, whose values are set as follows: $w_1=3.0$, $w_2=5.0$, $\theta_\alpha=3.0$, $\theta_\beta=50.0$, and $\theta_\gamma=3.0$. Here, all experimental configurations are trained with the same data. And we also select the best ones according to the validate set for performance evaluation.

Figure 6 shows the salient object detection results with above experimental settings on all test datasets in terms of wF_β. From Fig. 6, we can see that (1) E/E+D is better than R/R+D, since E has larger receptive field and can capture more global information about salient objects. (2) $E+R+D$ is better than $E+E+D$ and $R+R+D$, meaning that E/R is more suitable than R/E for global/local saliency estimation/refinement. (3) The generator G containing two components gets better performance than the one with only one component, for example, $E+R$ is better than E or R and $E+E+D$ is better than $E+D$ or $R+D$, meaning that the saliency refinement stage is important for performance improvement no matter using E or R. (4) R is better than E for saliency refinement, meaning that the proposed deep residual network R is more powerful than E for correcting the poor or wrong predictions by well leveraging the local contextual information of salient objects. (5) The performances when with D are better than those without D, meaning that the adversarial loss considered as a structured loss is able to help the models to better learn the structural information of salient objects. (6) Compared with using loss L_C or L_A alone, we can obtain large improvements by leveraging their combination in training. (7) $E+R+D$ gets better performance than $\underline{E+R+D}$ and the largest improvement is achieved on SED2 dataset where all images have two salient objects, suggesting that using our data augmentation technique can alleviate the effect of some biases (*e.g.*, single object and center bias), so as to increase the diversity of the training samples (*e.g.*, larger spatial distribution of salient objects and more salient objects), which makes the model be more powerful to tackle complex scenario, such as containing multiple salient objects. (8) $E+CRF$ is better than E, but still worse than $E+R$ and $E+R+D$, suggesting that our refinement model including the local saliency refiner R and adversarial learning is more effective than the CRF model for saliency refinement. We also calculate the runtime of CRF that is about 0.5 second on an image with size of 300×400 on CPU due to unavailable implementation of CRF on GPU, while our local saliency refiner R takes less than 0.05 second on GPU, suggesting that our refinement model is more efficient than CRF when using GPU. All of these results demonstrate that both cascaded CNNs design and adversarial learning are important for performance improvement and all components are essential for our method to achieve the best performance. The similar conclusions are obtained when using F_β, SSM and MAE. Due to the spatial limitation, only wF_β is presented here.

V. Conclusions

In this paper, we propose a novel end-to-end salient object detection model (CCAL) based on cascaded convolutional

neural networks and adversarial learning. A encoder-decoder network and a deep residual network constituting the cascaded CNNs are designed and play their roles in global saliency estimation and local saliency refinement, respectively. With the cascaded coarse-to-fine fashion, the performance of salient object detection can be boosted progressively. As a structured loss function, the adversarial loss introduced from discriminator is very useful to help CCAL to better learn the structural information of salient objects, and the experimental results illustrate its importance for performance improvement. The proposed method can produce accurate salient object detection results without any post-process. Experiments demonstrate that CCAL not only gets the state-of-the-art performance over eight benchmark datasets, but also achieves a speed about 17 fps on GPU.

ACKNOWLEDGMENT

This work was supported in part by the Natural Science Foundation of China under Grant 61672194, by the National Key R&D Program of China under Grant 2018YFC0832304, by the Distinguished Youth Science Foundation of Heilongjiang Province of China under Grant JC2018021, by the Shandong Provincial Natural Science Foundation, China under Grant ZR2016FM04, and by the Humanity and Social Science Youth foundation of Ministry of Education of China under Grant 14YJC760001.

REFERENCES

[1] C. Jung and C. Kim, "A unified spectral-domain approach for saliency detection and its application to automatic object segmentation," *IEEE Trans. Image Process.*, vol. 21, no. 3, pp. 1272–1283, 2012.

[2] 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.

[3] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2004, pp. 37–44.

[4] Q. Yan, L. Xu, J. Shi, and J. Jia, "Hierarchical saliency detection," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2013, pp. 1155–1162.

[5] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2014, pp. 2814–2821.

[6] J. Kim, D. Han, Y.-W. Tai, and J. Kim, "Salient region detection via high-dimensional color transform," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2014, pp. 883–890.

[7] Y. Qin, H. Lu, Y. Xu, and H. Wang, "Saliency detection via cellular automata," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2015, pp. 110–119.

[8] 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, 2015.

[9] J. Sun, H. Lu, and X. Liu, "Saliency region detection based on markov absorption probabilities," *IEEE Trans. Image Process.*, vol. 24, no. 5, pp. 1639–1649, 2015.

[10] C. Li, Y. Yuan, W. Cai, Y. Xia, and D. Dagan Feng, "Robust saliency detection via regularized random walks ranking," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2015, pp. 2710–2717.

[11] M.-M. Cheng, N. J. Mitra, X. Huang, P. H. Torr, and S.-M. Hu, "Global contrast based salient region detection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 3, pp. 569–582, 2015.

[12] Y. Tang, X. Wu, and W. Bu, "Saliency detection based on graph-structural agglomerative clustering," in *Proc. ACM Multimedia*, 2015, pp. 1083–1086.

[13] C. Scharfenberger, A. Wong, and D. A. Clausi, "Structure-guided statistical textural distinctiveness for salient region detection in natural images," *IEEE Trans. Image Process.*, vol. 24, no. 1, pp. 457–470, 2015.

[14] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2015, pp. 1884–1892.

[15] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2009, pp. 1597–1604.

[16] J. Zhang, S. Sclaroff, Z. Lin, X. Shen, B. Price, and R. Mech, "Minimum barrier salient object detection at 80 fps," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1404–1412.

[17] P. Jiang, N. Vasconcelos, and J. Peng, "Generic promotion of diffusion-based salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 217–225.

[18] W. Zou and N. Komodakis, "Harf: Hierarchy-associated rich features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 406–414.

[19] L. Wang, H. Lu, X. Ruan, and M.-H. Yang, "Deep networks for saliency detection via local estimation and global search," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2015, pp. 3183–3192.

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

[21] R. Zhao, W. Ouyang, H. Li, and X. Wang, "Saliency detection by multi-context deep learning," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2015, pp. 1265–1274.

[22] Y. Tian, J. Li, S. Yu, and T. Huang, "Learning complementary saliency priors for foreground object segmentation in complex scenes," *Int. J. Comput. Vis.*, vol. 111, no. 2, pp. 153–170, 2015.

[23] L. Wang, L. Wang, H. Lu, P. Zhang, and X. Ruan, "Saliency detection with recurrent fully convolutional networks," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 825–841.

[24] Y. Tang and X. Wu, "Saliency detection via combining region-level and pixel-level predictions with cnns," in *Proc. Eur. Conf. Comput. Vis.*, 2016, pp. 809–825.

[25] N. Liu and J. Han, "Dhsnet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2016, pp. 678–686.

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

[27] Y. Tang, X. Wu, and W. Bu, "Deeply-supervised recurrent convolutional neural network for saliency detection," in *Proc. ACM Multimedia*, 2016, pp. 397–401.

[28] J. Kuen, Z. Wang, and G. Wang, "Recurrent attentional networks for saliency detection," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2016, pp. 3668–3677.

[29] 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, 2015.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[31] S. Gidaris and N. Komodakis, "Object detection via a multi-region and semantic segmentation-aware cnn model," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1134–1142.

[32] L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, and X. Ruan, "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2017, pp. 136–145.

[33] Q. Hou, M. M. Cheng, X. Hu, A. Borji, Z. Tu, and P. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2017, pp. 5300–5309.

[34] C. Xia, J. Li, X. Chen, A. Zheng, and Y. Zhang, "What is and what is not a salient object? learning salient object detector by ensembling linear exemplar regressors," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit*, 2017, pp. 4321–4329.

[35] X. Chen, A. Zheng, J. Li, and F. Lu, "Look, perceive and segment: Finding the salient objects in images via two-stream fixation-semantic cnns," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 1050–1058.

[36] Z. Luo, A. Mishra, A. Achkar, J. Eichel, S. Li, and P.-M. Jodoin, "Non-local deep features for salient object detection," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2017, pp. 6609–6617.

[37] P. Zhang, D. Wang, H. Lu, H. Wang, and X. Ruan, "Amulet: Aggregating multi-level convolutional features for salient object detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 202–211.

[38] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 212–221.

[39] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 4039–4048.

[40] L. Ye, Z. Liu, L. Li, L. Shen, C. Bai, and Y. Wang, "Salient object segmentation via effective integration of saliency and objectness," *IEEE Trans. Multimedia*, vol. 19, no. 8, pp. 1742–1756, 2017.

[41] Z. Wang, D. Xiang, S. Hou, and F. Wu, "Background-driven salient object detection." *IEEE Trans. Multimedia*, vol. 19, no. 4, pp. 750–762, 2017.

[42] R. Quan, J. Han, D. Zhang, F. Nie, X. Qian, and X. Li, "Unsupervised salient object detection via inferring from imperfect saliency models," *IEEE Trans. Multimedia*, vol. 20, no. 5, pp. 1101–1112, 2018.

[43] H. Xiao, J. Feng, Y. Wei, M. Zhang, and S. Yan, "Deep salient object detection with dense connections and distraction diagnosis," *IEEE Trans. Multimedia*, vol. 20, no. 12, pp. 3239–3251, 2018.

[44] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, and A. Borji, "Detect globally, refine locally: A novel approach to saliency detection," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2018, pp. 3127–3135.

[45] L. Zhang, J. Dai, H. Lu, Y. He, and G. Wang, "A bi-directional message passing model for salient object detection," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2018, pp. 1741–1750.

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

[47] Y. Tang and X. Wu, "Text-independent writer identification via cnn features and joint bayesian," in *Proc. Int. Conf. Frontiers in Handwriting Recognit.*, 2016, pp. 566–571.

[48] J. Cai, Y. Tang, L. Lu, A. P. Harrison, K. Yan, J. Xiao, L. Yang, and R. M. Summers, "Accurate weakly-supervised deep lesion segmentation using large-scale clinical annotations: Slice-propagated 3d mask generation from 2d recist," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent*, 2018, pp. 396–404.

[49] Y. Tang and X. Wu, "Scene text detection using superpixel-based stroke feature transform and deep learning based region classification," *IEEE Trans. Multimedia*, vol. 20, no. 9, pp. 2276–2288, 2018.

[50] Y. Tang, A. P. Harrison, M. Bagheri, J. Xiao, and R. M. Summers, "Semi-automatic recist labeling on ct scans with cascaded convolutional neural networks," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent*, 2018, pp. 405–413.

[51] Y. Tang and X. Wu, "Scene text detection via edge cue and multi-features," in *Proc. Int. Conf. Frontiers in Handwriting Recognit.*, 2016, pp. 156–161.

[52] D. Jin, Z. Xu, Y. Tang, A. P. Harrison, and D. J. Mollura, "Ct-realistic lung nodule simulation from 3d conditional generative adversarial networks for robust lung segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent*, 2018, pp. 732–740.

[53] Y. Tang, J. Cai, L. Lu, A. P. Harrison, K. Yan, J. Xiao, L. Yang, and R. M. Summers, "Ct image enhancement using stacked generative adversarial networks and transfer learning for lesion segmentation improvement," in *Proc. Int. Conf. Mach. Learn. Med. Imag.*, 2018, pp. 46–54.

[54] M. Mirza and S. Osindero, "Conditional generative adversarial nets," *arXiv preprint arXiv:1411.1784*, 2014.

[55] J. Pan, C. Canton, K. McGuinness, N. E. O'Connor, J. Torres, E. Sayrol, and X. Giro-i Nieto, "Salgan: Visual saliency prediction with generative adversarial networks," *arXiv preprint arXiv:1701.01081*, 2017.

[56] H. Noh, S. Hong, and B. Han, "Learning deconvolution network for semantic segmentation," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1520–1528.

[57] Y. Tang and X. Wu, "Salient object detection with chained multi-scale fully convolutional network," in *Proc. ACM Multimedia*, 2017, pp. 618–626.

[58] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *Proc. Int. Conf. Med. Image Comput. Comput. Assist. Intervent*, 2015, pp. 234–241.

[59] Y. Tang and X. Wu, "Scene text detection and segmentation based on cascaded convolution neural networks," *IEEE Trans. Image Process.*, vol. 26, no. 3, pp. 1509–1520, 2017.

[60] K. He, X. Zhang, S. Ren, and J. Sun, "Delving deep into rectifiers: Surpassing human-level performance on imagenet classification," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1026–1034.

[61] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proc. Int. Conf. Mach. Learn.*, 2015, pp. 448–456.

[62] S. Gross and M. Wilber, "Training and investigating residual nets," in *online at http://torch.ch/blog/2016/02/04/resnets.html*, 2016.

[63] P. Krähenbühl and V. Koltun, "Efficient inference in fully connected crfs with gaussian edge potentials," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 109–117.

[64] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A. Aitken, A. Tejani, J. Totz, Z. Wang *et al.*, "Photo-realistic single image super-resolution using a generative adversarial network," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2017, pp. 4681–4690.

[65] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Proc. Adv. Neural Inf. Process. Syst.*, 2014, pp. 2672–2680.

[66] P. Isola, J. Y. Zhu, T. Zhou, and A. A. Efros, "Image-to-image translation with conditional adversarial networks," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2017, pp. 5967–5976.

[67] X. Mao, Q. Li, H. Xie, R. Y. Lau, and Z. Wang, "Multi-class generative adversarial networks with the l2 loss function," *arXiv preprint arXiv:1611.04076*, 2016.

[68] S. Xie and Z. Tu, "Holistically-nested edge detection," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 1395–1403.

[69] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin *et al.*, "Tensorflow: Large-scale machine learning on heterogeneous distributed systems," *arXiv preprint arXiv:1603.04467*, 2016.

[70] D. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[71] S. Alpert, M. Galun, A. Brandt, and R. Basri, "Image segmentation by probabilistic bottom-up aggregation and cue integration," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 2, pp. 315–327, 2012.

[72] 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. and Pattern Recognit.*, 2014, pp. 280–287.

[73] V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.W*, 2010, pp. 49–56.

[74] C. Yang, L. Zhang, H. Lu, X. Ruan, and M.-H. Yang, "Saliency detection via graph-based manifold ranking," in *Proc. IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2013, pp. 3166–3173.

[75] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2001, pp. 416–423.

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

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

Youbao Tang is a postdoc researcher in National Institutes of Health (NIH), US. He received the B.Sc., M.Sc., and Ph.D. degrees in Computer Science from Harbin Institute of Technology (HIT), Harbin, China, in 2009, 2011, and 2016, respectively. His research interests include image processing, computer vision, biometrics, and medical image analysis.

Xiangqian Wu (M'06, SM'17) received the B.Sc., M.Sc., and Ph.D. degrees in computer science from the Harbin Institute of Technology (HIT), Harbin, China, in 1997, 1999, and 2004, respectively. Prof. Wu has worked as a lecturer (2004-2006), associate professor (2006-2009) and professor (2009-present) at the School of Computer Science and Technology, HIT. He has published one book and more than 100 papers in international journals and conferences. His current research interests include computer vision, pattern recognition, biometrics and medical image analysis, etc.