Neurocomputing 381 (2020) 130–140

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Lightweight adversarial network for salient object detection

Lili Huang[a], Guanbin Li[a,*], Ya Li[b], Liang Lin[a]

[a] School of Data and Computer Science, Sun Yat-sen University, Guangzhou 510006, China
[b] Guangzhou University, Guangzhou 510182, China

ARTICLE INFO

Article history:
Received 13 February 2019
Revised 18 July 2019
Accepted 1 September 2019
Available online 11 November 2019

Communicated by Dr. Liquan Shen

MSC:
00-01
99-00

Keywords:
Salient object detection
Lightweight model
Deep learning
Adversarial training
Multi-scale feature

ABSTRACT

Recent advance on salient object detection benefits mostly from the revival of Convolutional Neural Networks (CNNs). However, with these CNN based models, the predicted saliency map is usually incomplete, that is, spatially inconsistent with the corresponding ground truth, because of the inherent complexity of the object and the inaccuracy of object boundary detection resulted from regular convolution and pooling operations. Besides, the breakthrough on saliency detection accuracy of current state-of-the-art deep models comes at the expense of high computational cost, which contradicts its role as a pretreatment procedure for other computer vision tasks. To alleviate these issues, we propose a lightweight adversarial network for salient object detection, which simultaneously improves the accuracy and efficiency by enforcing higher-order spatial consistency via adversarial training and lowering the computational cost through lightweight bottleneck blocks, respectively. Moreover, multi-scale contrast module is utilized to sufficiently capture contrast prior for visual saliency reasoning. Comprehensive experiments demonstrate that our method is superior to the state-of-the-art works on salient object detection in both accuracy and efficiency.

1. Introduction

Salient Object Detection aims at identifying the most attention-drawing objects in an image and then pixel-wise segmenting these objects with binary labels, as illustrated in Fig. 1. As with the human subconscious, salient object detection generally services as the pretreatment procedure for many other computer vision tasks, such as visual tracking [1], content-aware image editing [2,3], and weakly supervised semantic segmentation [4,5]. Therefore, it is expected to handle with saliency inferring accurately and efficiently.

Recent years, with the revival of the convolutional neural networks (CNNs) [6], various CNN based models are explored for salient object detection. However, when using these models, the predicted saliency map is usually incomplete, that is, spatially inconsistent with the corresponding ground truth, because of the inherent complexity of the object and the inaccuracy of object boundary detection resulted from regular convolution and pooling operations. On the contrary, the discriminator distinguishes the authenticity of the input image and further leads to a more complete saliency detection result, which inspires us to introduce adversarial training into the visual saliency learning. At present, conditional Markov random fields (CRFs) [7–11] are the most commonly used post-processing methods, which directly integrate pairwise and specific higher-order potentials into the CNN based models. Despite CRFs can reinforce spatial contiguity in the predicted saliency maps, they are limited in parameter number of high-order potentials. Instead, in this paper, we utilize adversarial training to resolve this problem while improving the performance on accuracy for salient object detection.

On the other hand, the breakthrough on saliency detection accuracy of current state-of-the-art deep models [12–22] comes at the expense of high computational cost, which contradicts its role as a pretreatment procedure for other computer vision tasks and thus greatly limits its pervasive application on embedded and mobile devices. To improve the performance on efficiency, we intend to explore a lightweight framework. Albeit varieties of researches delve into lightweight model designs, such as knowledge distillation [23–25] and network pruning [26–31], connectivity learning [32,33] and hyper-parameter optimization [34,35], they are complicated and lacked of universality. For example, network pruning approaches necessitate pre-trained large models to obtain smaller models with the comparable performance. Moreover, these models cannot be directly applied into salient object detection, because they are not tailor-designed for capturing subtle visual contrast, which is the most significant factor [36,37] for accuracy improvement.

Numerous strategies are used to explore local or global contrast cues for visual saliency detection. Previous works capture

* Corresponding author.
E-mail address: liguanbin@mail.sysu.edu.cn (G. Li).

https://doi.org/10.1016/j.neucom.2019.09.100
0925-2312/

Fig. 1. Examples of salient object detection. The first row shows the input RGB images, and the second row presents the ground truth for salient object detection. Best viewed in color.

visual contrast through sophisticated hand-crafted low-level features, such as color, intensity and texture. As illustrated in the first two columns of Fig. 1, visual contrast of the two examples can be represented by low-level color features. On the contrary, salient objects in the last two columns hardly stand out from the background, because they have the similar appearance. Recently, CNN based models have been employed to obtain high-level semantic features, which are more robust than hand-crafted ones, achieving better results than early attempts. However, most of the current deep methods still lack efficient strategy to exploit multi-scale global contrast context. As discussed in [38], the amount of available context information depends on the size of receptive field in the deep neural network, however, the empirical receptive field of CNN is much smaller than the theoretical one, especially on high-level layers. Consequently, the networks cannot sufficiently model global context prior.

To alleviate the aforementioned issues, this work, inspired by the generative adversarial networks (GANs) [39,40] and lightweight models [41,42], proposes a Lightweight Adversarial Network (LANet) for salient object detection, which simultaneously improves accuracy and efficiency by enforcing higher-order spatial consistency through adversarial training and lowering the computational cost via lightweight bottleneck blocks, respectively. LANet utilizes encoder-decoder architecture based saliency predictor to generate the saliency map of an input image, where multi-scale contrast module is used to encode rich contextual information for visual saliency reasoning. During the training phase, this network is initially trained with a saliency loss over the predicted saliency maps. Afterwards, the model is finetuned with an adversarial network trained to solve a binary classification task between the saliency maps predicted by LANet and the ground-truth ones, meaning that an adversarial loss is incorporated into the visual saliency learning. It is worth mentioning that lightweight bottleneck blocks, instead of regular convolutions, are utilized in both saliency predictor and adversarial network to learn features, which has been proven very efficient in [41].

In summary, this paper has the following contributions:

- We propose an accurate and efficient network, i.e. LANet, for salient object detection, which utilizes lightweight linear bottleneck blocks to construct both saliency predictor and adversarial network, resulting in tremendous reduction in the computational cost to ensure the efficiency.
- We incorporate adversarial training into saliency prediction model to improve performance via enforcing long-range spatial saliency contiguity, while consuming no additional computation cost at the test phase. Multi-scale contrast module is also used to further improve accuracy via encoding rich contrast cues.
- This work presents comprehensive experiments on the trade-off of accuracy and efficiency. Experimental result demonstrates that our model significantly outperforms the state-of-the-art works on salient object detection.

2. Related work

Salient object detection. Over the past two decades, a lot of salient object detection approaches have been developed. Compared with traditional methods [43–46] that use hand-crafted features, recently emerged CNN-based methods have broken almost all the previous state-of-the-art records in nearly every sub-field of computer vision, including salient object detection. Here, we mainly focus on introducing the deep learning based methods on saliency detection. The rapidly sprung up deep models can be separated into two categories: patch based multi-stage approaches and end-to-end FCN-based approaches. The patch based approaches [12,13,47] first partition an image into patches and treat each patches as independent samples for training and testing, which are inefficient due to redundancy of overlapped patches. The FCN-based approaches [16,17,48,49] overcome the above deficiency. They take the whole image as input and train the model in an end-to-end way. Some approaches [16,48] capture high-level contrast by multi-scale features, which is extracted from extra convolution layers. Recently, Hou et al. [17] proposed a deeply supervised salient (DSS) model through adding connections from high-level features to low-level features based on the holistically-nested edge detector [50] architecture. Although these end-to-end networks improve accuracy and become a fundamental architecture for saliency detection, the necessity of high performance computing resources is still a challenging problem for applying them on embedded and mobile devices.

Adversarial learning. Generative adversarial network [39], composed of generative and adversarial networks, is an adversarial method for learning deep generative models. It is originally designed for image generation [39,51,52], where the generative network is used to produce fake images, while the adversarial network aims at distinguishing between the real images and fake ones

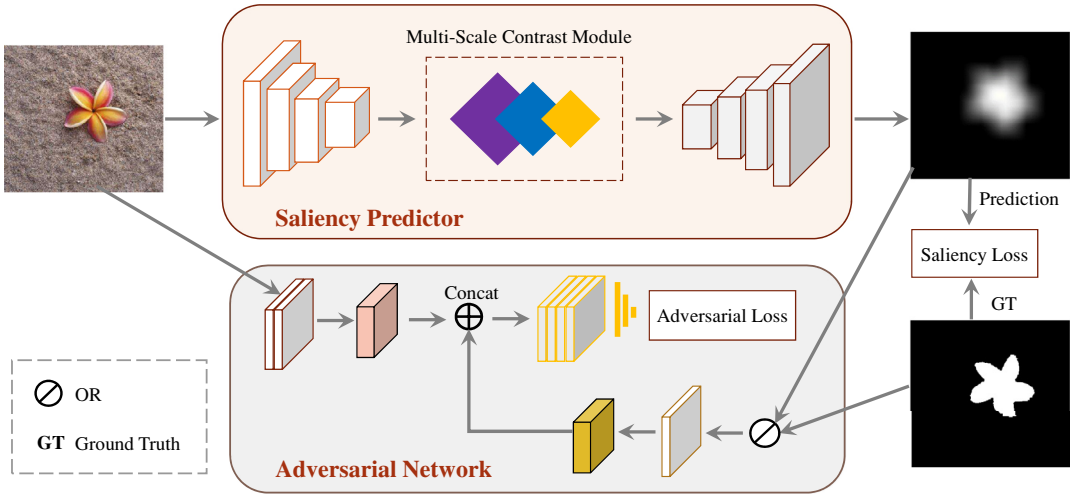


Fig. 2. Overview of our proposed lightweight adversarial network. Lightweight bottleneck blocks, instead of regular convolutions, are applied to both saliency predictor and adversarial network. Best viewed in color.

generated by the generative network. Later, GAN is repurposed for other research fields. For instance, image-conditioned GAN [53] is proposed for super-resolution, perceptual GAN [54] is used to handle the issue of small object detection, SalGAN [55] aims at generating saliency results of given images. Recently, some works [56,57] extend GAN to salient object detection.Pan et al. [56] utilize the features of GAN to deal with image salient object detection. Zhu et al. [57] model multi-scale adversarial feature to improve performance. In this paper, we present a lightweight adversarial network for salient object detection, which incorporates the high-level adversarial loss into the conventional saliency prediction during network training.

Lightweight deep model. Besides accuracy, efficiency is another important consideration in many real applications. This motivates a series of deep models towards lightweight architecture design. Early works [58–61] optimize networks through manual tuning parameters. Subsequent works [26,28,30–32] capitalize on connectivity learning, network pruning, and hyper-parameter optimization to investigate new network architecture. Recent years, newly designed internal convolutional blocks, such as group convolution and depth-wise convolution, are utilized to current lightweight deep models [41,42,62–64]. However, none of them are designed for visual saliency inferring. In this paper, we propose a lightweight and effective deep framework for salient object detection.

3. Lightweight adversarial network

As illustrated in Fig. 2, our designed LANet consists of saliency predictor and adversarial network, both of which are built upon lightweight bottleneck blocks, instead of regular convolutions, resulting in an efficient end-to-end solution. Saliency predictor involves feature extractor, multi-scale contrast module and decoder. It is used to generate saliency maps. Specifically, given an image, we adopt feature extractor, i.e. tailored VGG, to produce the initial feature maps, and then feed them into the multi-scale contrast module to produce the final encoder feature maps. Afterward, the resulting feature maps are upsampled through a decoder to yield final saliency prediction. During the training phase, this network is initially trained with a saliency loss over the predicted saliency maps. Then the model is finetuned with an adversarial network trained to solve a binary classification task between the saliency maps predicted by LANet and the corresponding ground truth. In this section, we provide details of the proposed LANet.

3.1. Lightweight bottleneck block

Since lightweight bottleneck block is one of the fundamental units of each component in our proposed LANet, we first give its detail description in this section. As illustrated in (a) of Fig. 3, the lightweight bottleneck blocks are characterized by depth-wise separable convolutions [65], linear bottlenecks, and inverted residuals. In particular, a depth-wise separable convolution is sequentially decomposed into a depth-wise convolution and a point-wise convolution, i.e. 1×1 convolution. The depth-wise convolutions filter each input channel with different convolution kernels, while the succeeding point-wise convolutions further fuse the resultant features across all channels to produce new feature representations. It can be intuitively observed that the lightweight bottleneck block first expands the input to higher dimension with a point-wise convolution, then filters the high-dimensional result with a depth-wise convolution, and finally reduces dimension of resultant features via another point-wise convolution.

The main contribution of the depth-wise separable convolution is tremendous reduction in the computational cost. To demonstrate this advantage, we compare computational costs of the depth-wise separable convolution and the regular convolution. Given an input tensor $T_i \in \mathbb{R}^{C_i \times H_i \times W_i}$, we convolve it with kernel size $K \times K$, resulting feature map $T_i' \in \mathbb{R}^{C_j \times H_i \times W_i}$. Thus, the corresponding computational cost M of a depth-wise separable convolution is:

$$\begin{aligned} M &= (K^2 \times C_i \times H_i \times W_i) + (C_i \times C_j \times H_i \times W_i) \\ &= (K^2 + C_j) \times C_i \times H_i \times W_i, \end{aligned} \tag{1}$$

while the computational cost R of a regular convolution operation is:

$$R = K^2 \times C_j \times C_i \times H_i \times W_i. \tag{2}$$

Therefore, the computation cost of depth-wise separable convolutions is almost K^2, i.e. $K^2 C_j/(K^2 + C_j)$, times smaller than that of regular convolutions.

The inverted residual in a lightweight bottleneck block is used to speed up the procedures of training and inference. As illustrated in (a) of Fig. 3, it establishes shortcut link between the input layers and the layers after dimensionality reduction, which happens only if the stride is 1. As can be seen, the difference between the inverted residual and the classical residual in (b) is that: the inverted residual in (a) links the thinner cubes (feature maps) while the classical residual in (b) connects the thicker cubes. Moreover,

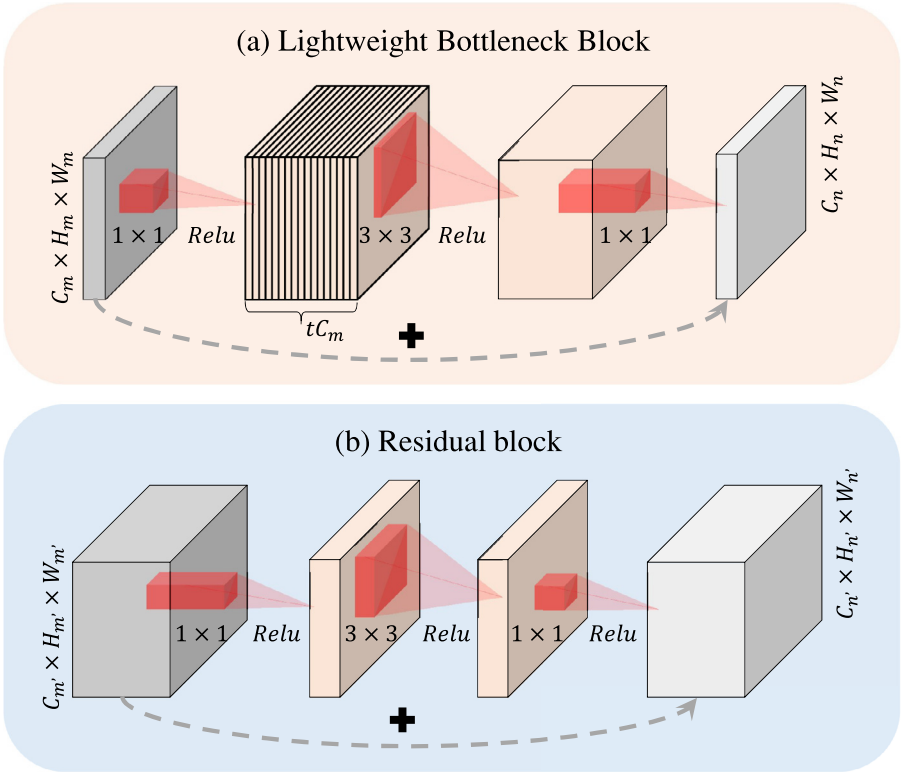


Fig. 3. Comparison between the lightweight bottleneck block [41] and residual block [61]. + indicates an addition operation which happens only if the stride is 1, while the dashed arrows in (a) and (b) represent the inverted residual connection and residual connection, respectively. The thickness of each cube (i.e., feature maps), denotes its relative number of channels. t is the expansion rate.

to hinder useful information from being destroyed, the bottleneck block gets rid of the non-linear activation in its last layer.

3.2. Adversarial training

As discussed in aforementioned sections, when using CNN based models, the predicted saliency map is usually incomplete, that is, spatially inconsistent with the corresponding ground truth. Existed post-processing methods, e.g. CRFs, are limited in parameter number of high-order potentials. Instead, we utilize adversarial training, incorporating an adversarial loss into saliency predictor, to enforce higher-order spatial consistency. Specifically, we take into account both saliency loss ℓ_s and adversarial loss ℓ_a to train LANet. For the sake of presentation, binary cross-entropy loss used in the two terms is denoted as

$$\ell_{bce}(\hat{z}, z) = -[z \ln \hat{z} + (1 - z) \ln(1 - \hat{z})], \tag{3}$$

where $z \in \{0, 1\}$ is binary indicator.

Saliency loss is a standard binary cross-entropy term that enforces the saliency model to independently predict the accurate saliency at each pixel location, making higher-order spatial consistency unachievable. Given a data set of N training images $x^k \in \mathbb{R}^{H \times W \times 3}$ and a corresponding ground-truth saliency map $y^k \in \{0, 1\}^{H \times W}$, we use $\hat{y} = S(x)$ to denote the predicted saliency map. Then, saliency loss can be computed as

$$\ell_s(\hat{y}, y) = \sum_{i=1}^{H \times W} \ell_{bce}(\hat{y}_i, y_i) = -\sum_{i=1}^{H \times W} [y_i \ln \hat{y}_i + (1 - y_i) \ln(1 - \hat{y}_i)]. \tag{4}$$

Since the adversarial network handles binary classification problems, resulting in much larger field-of-view than that of saliency predictor, the adversarial training can enforce higher-order spatial consistency. We denote $A(x, y) \in [0, 1]$ as the probability that adversarial network predicts y to be the ground-truth saliency map of x, as opposed to being the output of the saliency predictor $S(\bullet)$. Thus adversarial loss is also represented by the binary cross-entropy loss:

$$\ell_a(x, y) = \ell_{bce}(A(x, y), 1) + \ell_{bce}(A(x, S(x)), 0). \tag{5}$$

Training the saliency predictor. Given the adversarial network, the training of the saliency predictor aims at minimizing the saliency loss, while simultaneously degrading the performance of the adversarial network. This encourages the saliency predictor to produce saliency maps that are hard to distinguish from ground-truth ones for the adversarial network. As proven in work [39], it propagates stronger gradient to maximize $\ell_{bce}(A(x^k, S(x^k)), 1)$ than to minimize $\ell_{bce}(A(x^k, S(x^k)), 0)$, thus the loss function relevant to the saliency predictor is

$$\frac{1}{N} \sum_{k=1}^{N} [\ell_s(S(x^k), y^k) + \lambda \ell_{bce}(A(x^k, S(x^k)), 1)]. \tag{6}$$

Training the Adversarial Network. Since adversarial loss ℓ_s only depends on the adversarial network, training the adversarial network is equivalent to minimizing the following binary classification loss

$$\frac{1}{N} \sum_{k=1}^{N} \ell_a(x, y) = \frac{1}{N} \sum_{k=1}^{N} \ell_{bce}(A(x^k, y^k), 1) + \ell_{bce}(A(x^k, S(x^k)), 0). \tag{7}$$

It is worth mentioning that incorporating the adversarial loss does not consume additional computation resources at the test phase. Lightweight bottleneck blocks, fundamental units of adversarial network, also make training more efficient. The architectures of

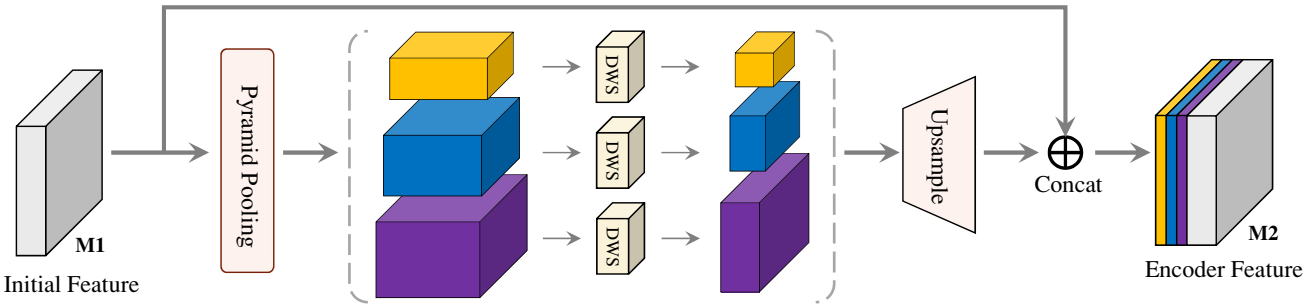


Fig. 4. Illustration of our proposed multi-scale contrast module. *DWS* denotes depth-wise separable convolutions.

both saliency predictor and adversarial network will be depicted in the Section 3.4 below.

3.3. Multi-Scale contrast module

As discussed in the aforementioned section, our proposed LANet, built upon the bottleneck blocks, insufficiently incorporates the crucial global contrast prior, due to the substantial loss of the empirical receptive field in both vanilla CNNs and lightweight bottleneck blocks. To solve this problem, we propose a multi-scale contrast module to properly incorporate both global and local contextual prior for accurate salient object detection.

The spatial pyramid average pooling can utilize spatial statistics to effectively describe the whole image, which has been proven in [66,67]. In this work, we further extend pyramid average pooling to multi-scale visual contrast representation for salient object detection. As illustrated in Fig. 4, the multi-scale contrast module contains pyramid average pooling, depth-wise separable convolutions (i.e. depth-wise convolutions coupled with point-wise convolutions), and upsampling operation. To separately extract local and global visual contrast information from the initial features $M1$, we take into consideration three different-level size (i.e., small, median and large) while performing pyramid average pooling. The large-size pooling, with the filter size of 32×32, is the coarsest global pooling to generate a single bin output. The following other size pooling, with the filter sizes of 8×8, 2×2, respectively partitions the feature map into different subregions, yielding pooled representation for each corresponding sub-region with bin sizes of 4×4, 16×16. The resultant feature maps are respectively fed into a depth-wise separable convolution layer to reduce their channel number to 1/3 of the original one. Afterwards, we expand the different-size output features to the size of $M1$ by bilinear interpolation and concatenate them with initial features $M1$ to yield the final encoder features $M2$. In particular, we adopt the depth-wise separable convolution, rather than the 1×1 convolution to reduce the dimensionality, because the former can eliminate redundant information while retaining the more expressive features. The experiment in Section 4.3 shows that the depth-wise separable convolution boosts accuracy by about 0.6% with negligible computation cost.

3.4. Network architecture

Architecture of Saliency Predictor. In the saliency predictor, the feature extractor is composed of convolutional layers of VGG16 [59], where regular convolutions are replaced with lightweight bottleneck blocks to reduce model parameters, while the decoder produces saliency maps by upsampling with parameter-free bilinear interpolation by a factor of 16.

Architecture of Adversarial Network. The architecture of the adversarial network is illustrated in Fig. 5. It takes as input an RGB image, and the corresponding saliency map, both of which are resized to 224×224 before fed into the network. The saliency map is either the ground truth, or generated by the saliency predictor. As suggested in [68], feature maps generated from two separate branches in Fig. 5 have the same number of channels.

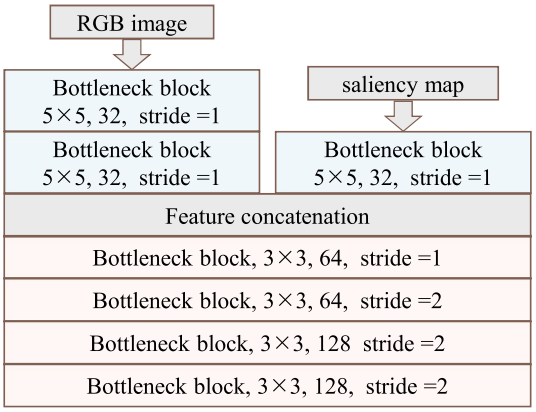


Fig. 5. Architecture of the adversarial network.

4. Experimental results

4.1. Experimental setup

Datasets. We evaluate the performance of our method on six widely used visual saliency datasets, including MSRA-B [69], HKU-IS [12], ECSSD [70], DUTOMRON [71], SOD [72,73] and PASCAL-S [74], all of which are available online. MSRA-B [69] comprises 5000 images with various image contents, most of which have only one coarsely annotated salient object. On contrast, images in HKU-IS [12] have low contrast and multiple salient objects. ECSSD [70] acquires 1000 natural images with meaningful semantics but complex structure from the Internet. Another large challenged dataset is DUTOMRON [71], comprising 5168 images, most of which are of multiple salient objects in complicated and cluttered backgrounds. SOD [72,73] comprises 300 images, each of which also possess multiple salient objects. PASCAL-S [74] comprises 850 images with the ground-truth masks annotated by 12 subjects. It was built upon the validation set of the PASCAL VOC 2010 segmentation challenge. Many images in this dataset face the challenges that: multiple salient objects are under occlusion or low contrast. To obtain a fair comparison with other methods, as done in [12,49,75], we combine the training sets of both the MSRA-B dataset [69] and the HKU-IS dataset [12] as our training set for salient object detection. The validation sets in the above two datasets are also combined as our validation set. Then we directly applied the trained model to test over every dataset.

Table 1
Comparison of the size and the computational cost between different networks for salient object detection. See Section 4.2 for details.

Methods	Params (M)	MADD (G)	GPU(s)	CPU(s)
MC	116.56	194.95	2.949	71.545
MDF	56.87	21.68	29.141	750.768
DS	134.27	180.88	0.191	4.652
RFCN	137.70	181.65	4.863	40.259
DCL	66.25	447.91	0.490	7.692
DSS	62.24	250.73	0.737	7.221
SRM	43.74	41.92	0.113	2.135
NLDF	35.58	279.13	0.254	7.235
UCF	23.99	80.18	0.243	15.119
C2S	137.05	107.02	0.160	20.039
RAS	80.93	230.19	0.058	6.806
LANet	**2.09**	**5.50**	**0.023**	**0.340**

Evaluation criteria. We evaluate the performance on both accuracy and efficiency. Specifically, the accuracy is evaluated using F-measure, mean absolute error (MAE) and precision-recall (PR) curves. The predicted saliency maps are converted to a binary masks using a threshold. The precision and recall is calculated by comparing the binary mask against the ground truth. The PR curve is produced by averaging precision and recall over saliency maps of a given dataset. The F-measure is formulated as

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

where β^2 is set as 0.3 to highlight the importance of the precision as suggested in [17,76]. The maximum F-measure (maxF) calculated from the PR curve is reported. MAE [45] measures the numerical distance between the ground truth G and an predicted saliency map M in a pixel-wise manner,

$$MAE = \frac{1}{W \times H}\sum_{i=1}^{W}\sum_{j=1}^{H}|M(i,j) - G(i,j)|, \tag{9}$$

where W and H denote the width and height of the saliency map, $M(i, j)$ denotes the saliency value of the pixel at (i, j) and the same for $G(i, j)$. On the other hand, the efficiency is measured by multiply-adds (MADD), actual latency, and the number of parameters as in [41].

Implementation details. Our LANet is implemented on the tensorflow [77], a flexible open source architecture with strong support for deep learning. We take the saliency predictor without multi-scale contrast module as a baseline model. Then, the baseline model integrating with adversarial network and the multi-scale contrast module acts as our final model for still image salient object detection when comparing with other benchmarks and performing the ablation study. During training and testing, the images are all resized to 512*512 through zero padding before fed into the saliency predictor. We train our framework using RMSPropOptimizer with both decay and momentum set to 0.95. The learning rate is initially set to 0.045 and decayed by 0.9 per epoch. Batch normalization is adopted after each convolution and before activation. The expansion rates in the bottleneck blocks are all set to 6 as in [41]. Both saliency loss and adversarial loss functions are formulated as binary cross-entropy loss. We use our validation set to search an optimal hyper parameter λ. The adversarial network is trained using $\lambda = 15$. Experiments are performed on a desktop with a GeForce GTX TITAN Black GPU and a 3.60 GHz Intel processor. The batch size is set to 6 in our experiment.

Table 2
Quantitative performance comparison between different saliency methods using maximum F-measure (higher is better) and MAE (lower is better) on six public datasets. The best three results on each dataset are shown in **red**, **blue**, and **green**, respectively. See Section 4.2 for details.

	MSRA-B		HKU-IS		ECSSD		DUTOMRON		SOD		PASCAL-S	
Method	maxF	MAE	maxF	MAE	maxF	MAE	maxF	MAE	maxF	MAE	maxF	MAE
MC	0.872	0.062	0.781	0.098	0.822	0.107	0.701	0.089	0.708	0.184	0.721	0.147
MDF	0.885	0.104	0.860	0.129	0.833	0.108	0.677	0.095	0.785	0.155	0.764	0.145
DS	0.856	0.061	0.808	0.071	0.810	0.160	**0.765**	0.070	0.781	0.150	0.818	0.170
RFCN	0.926	0.062	0.895	0.079	0.898	0.097	0.747	0.072	0.805	0.161	0.827	0.118
DCL	0.916	0.047	0.892	0.054	0.898	0.071	0.733	0.084	**0.832**	0.126	0.822	0.108
DSS	0.927	**0.028**	**0.913**	**0.039**	**0.915**	**0.052**	**0.760**	0.072	**0.842**	**0.118**	0.830	**0.080**
FSN	-	-	0.895	**0.044**	**0.910**	**0.053**	0.741	0.073	0.781	0.127	0.827	0.095
SRM	0.910	0.041	0.892	0.046	**0.910**	0.056	0.707	**0.069**	0.792	0.132	0.783	0.127
NLDF	0.910	0.044	0.902	0.048	0.905	0.063	0.753	0.080	0.808	0.130	**0.831**	0.112
AMU	**0.928**	**0.025**	**0.918**	0.052	0.889	0.059	0.733	0.097	0.773	0.145	**0.834**	0.103
UCF	0.921	0.031	0.905	0.074	0.868	0.078	0.713	0.132	0.776	0.169	0.771	0.128
PAGR	-	-	0.897	0.048	0.904	0.061	-	-	-	-	0.815	0.094
C2S	0.899	0.048	0.887	0.046	0.902	0.054	0.731	0.080	0.786	**0.124**	**0.834**	**0.082**
HCA	0.857	0.077	0.784	0.104	0.814	0.111	0.637	0.152	-	-	0.765	0.194
RAS	**0.931**	0.032	0.900	0.045	0.908	0.056	0.758	**0.068**	0.809	**0.124**	0.804	0.105
LANet	**0.929**	**0.023**	**0.922**	**0.031**	**0.917**	**0.049**	**0.814**	**0.061**	**0.851**	**0.099**	**0.858**	**0.076**

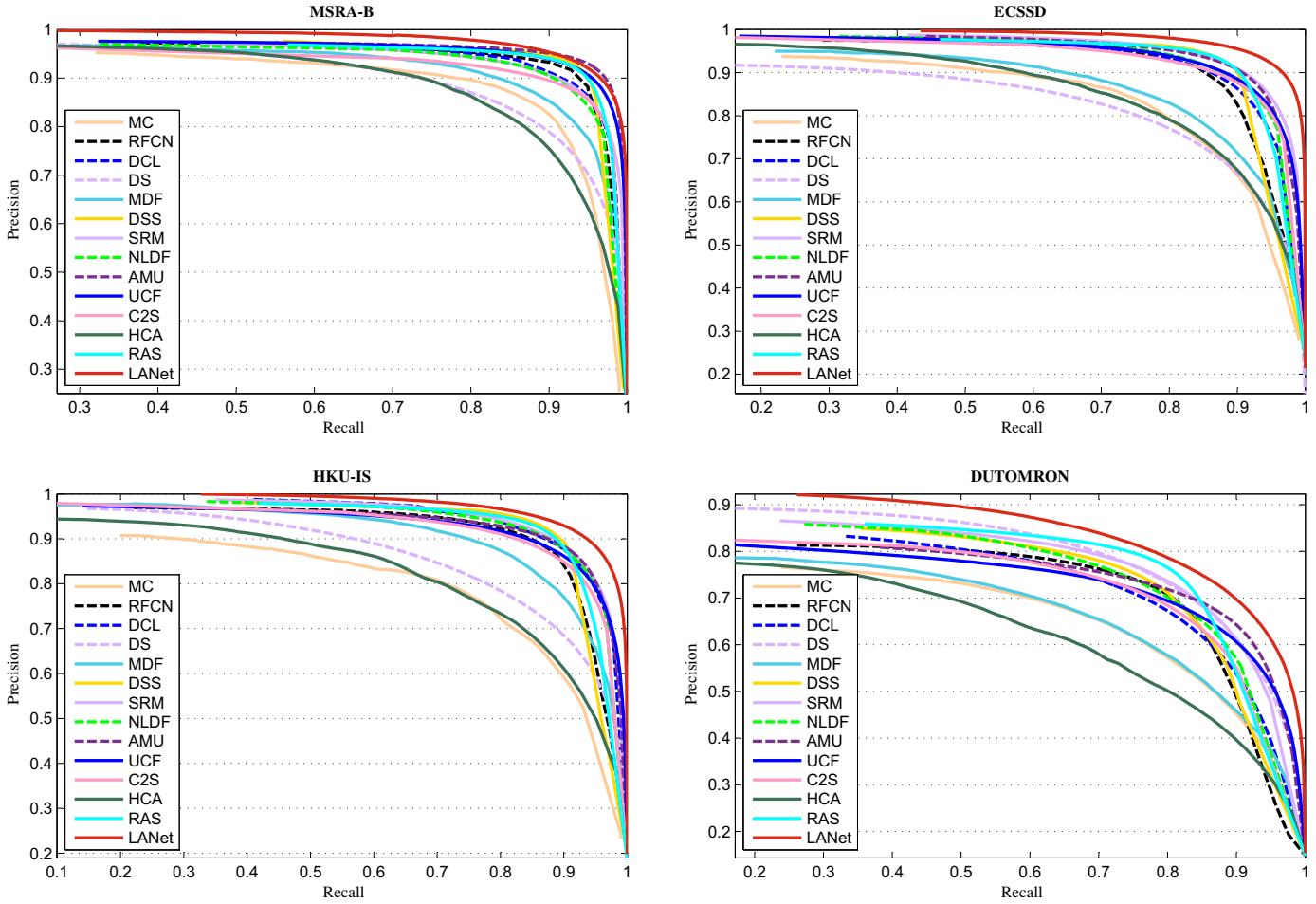


Fig. 6. Comparison of precision-recall curves on four benchmark datasets: MSRA-B [69], ECSSD [70], HKU-IS [12], DUTOMRON [71]. Our LANet consistently outperforms other methods across all the testing datasets. Best viewed in color. See Section 4.2 for details.

4.2. Comparison with the state of the art

We compare our LANet against 15 state-of-the-art salient object detection methods, including MC [13], MDF [12], DS [15], RFCN [14], DCL [16], DSS [17], FSN [78], SRM [79], NLDF [80], AMU [81], UCF [82], PAGR [83], C2S [84], HCA [85], and RAS [86]. For fair comparison, we use either the implementations or the saliency maps provided by the authors.

Table 1 show quantitative comparisons of the model size, the computational cost and running time between aforementioned 16 different models. As can be seen, benefiting from the lightweight bottleneck blocks, our proposed LANet yields the dramatic performance on efficiency: fewest parameters, lowest computational complexity, and least running time. Parameters of LANet is only 8.7% of that in the second fewest model UCF [12]. The computational cost of LANet is 5.50G multiply-adds, which is the lowest among comparison models. What's more, under lower-configured hardware, the running time of our LANet is 0.023 s on GPU and 0.340 s on CPU, which indicates that LANet is well suited for applications on the mobile and embedded platforms, and even without GPU, it also meets the real-time requirement.

We also provide quantitative evaluation of the performance on accuracy among different models. F-measure, MAE and PR curves are used for the evaluation on accuracy. Table 2 lists out the comparison results of F-measure and MAE. As can be observed, our approach significantly outperforms the competing methods both in terms of F-measure and MAE on six benchmark datasets. In particular, our LANet shows a significantly improved F-measure compared to the second best method, DS, for the DUTOMRON dataset (0.814 vs 0.765), which is one of the most challenging benchmarks. This clearly proves the superior performance of LANet in complex scenes. On the other hand, our LANet achieves a higher PR curve than all the other models, as shown in Fig. 6. Overall, our proposed model obtains the highest maximum F-measure and the lowest MAE on all the six datasets at a cost of least memory resource, due to the refinement effect of adversarial loss and multi-scale contrast features.

From another perspective, we present visual comparisons in Fig. 7. As can be seen, our predicted saliency is very close to the ground truth. In particular, our LANet yields considerable accurate saliency maps in various challenging cases, e.g., low contrast between saliency and background, multiple disconnected salient objects, and multi-scale salient objects. Moreover, our model provides more fine local details, such as shape and sharpness of saliency boundaries beside right salient region. To sum up, on account of the adversarial training and multi-scale contrast module, our LANet outperforms competing models in both accuracy and efficiency.

4.3. Ablation studies

Since adversarial training and multi-scale contrast module are two important components for our proposed LANet, we will demonstrate their effectiveness and necessity in this section. For the convenience of comparison studies, we denote the baseline

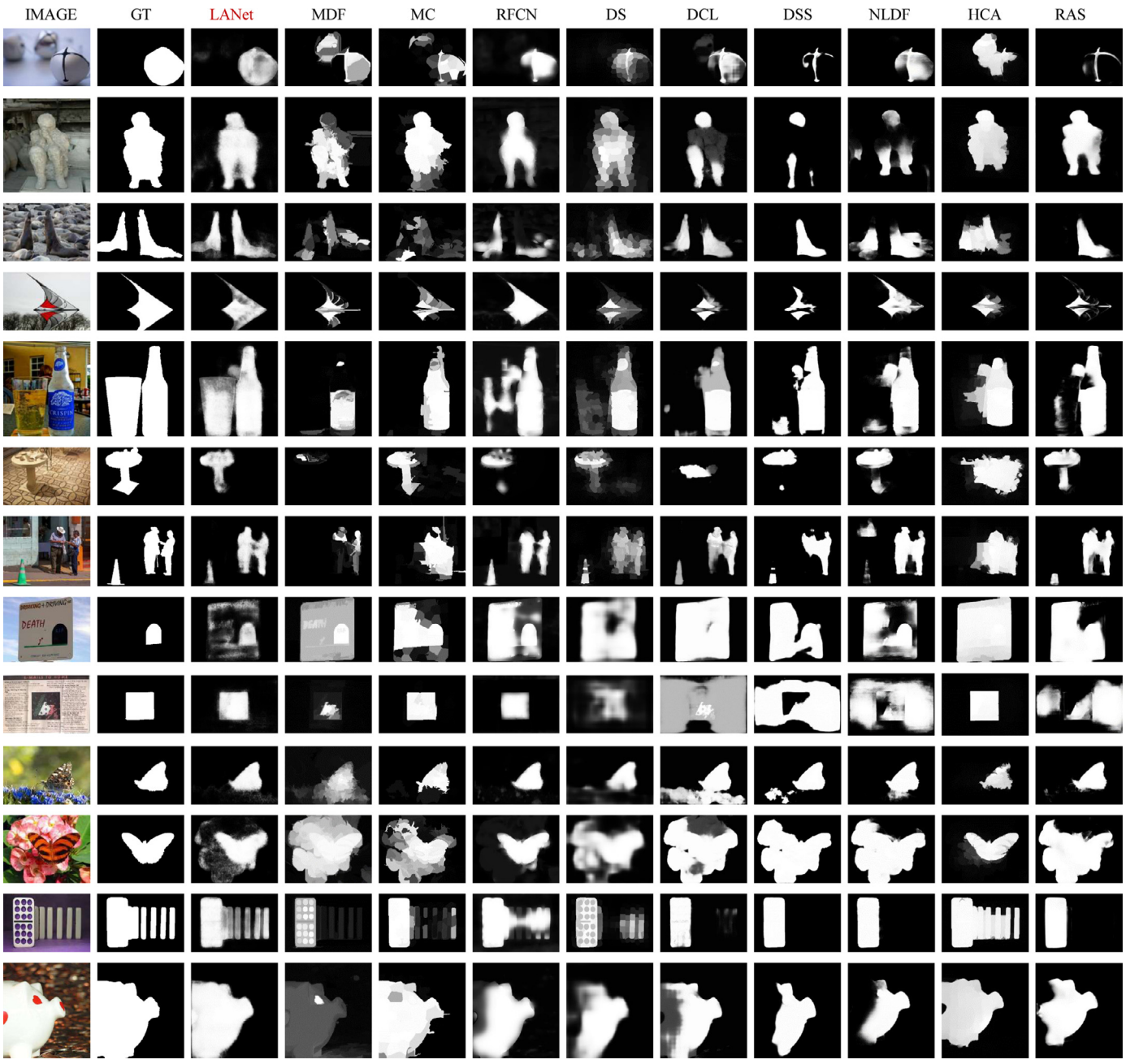


Fig. 7. Visual comparison of saliency maps generated from state-of-the-art methods, including our LANet. The ground truth (GT) is shown in the last column. Our model consistently produces saliency maps closest to the ground truth. Best viewed in color. See Section 4.2 for details.

model as BS, while the baseline model integrated with adversarial training or multi-scale contrast module are respectively abbreviated as AN, MSC. LANet is our final model for salient object detection.

Effectiveness of adversarial training. We first present accuracy performance comparison of the saliency produced using our LANet with and without adversarial training in the MSRA-B test set, i.e. LANet and MSC. As can be observed in Fig. 8, due to better enforcing long-range spatial contiguity, adversarial training method significantly improves performance on accuracy, which is in accordance with the comparison between the baseline model with and without adversarial training, i.e. AN and BS. On the other hand, we also compare the performance on efficiency, i.e. the model size, the computational cost and running time, between the models with and without adversarial training. As listed in Table 3, the performance of LANet and AN in all the metrics are the same as that in MSC and BS respectively. These results prove that it does not consume additional computation resources during the test phase to incorporate the adversarial loss into our model. Moreover, lightweight bottleneck blocks make it more memory-efficient to train the whole network.

Table 3
Comparison of the size and the computational cost between different design options. See Section 4.3 for details.

Methods	Params (M)	MADD (G)	GPU(s)	CPU(s)
BS	1.83	4.14	0.022	0.333
MSC	2.09	5.50	0.023	0.340
AN	1.83	4.14	0.022	0.333
LANet	**2.09**	**5.50**	**0.023**	**0.340**

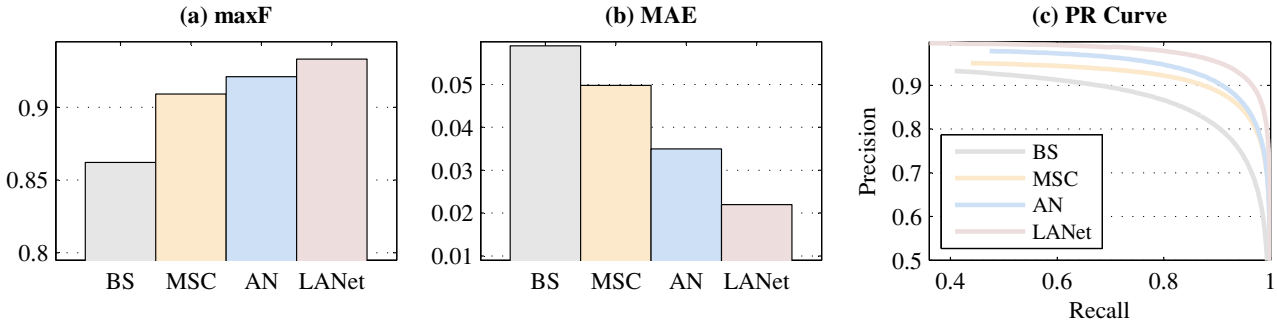


Fig. 8. Performance comparison between different design options. Best viewed in color. See Section 4.3 for details.

To further demonstrate adversarial training can better alleviate the problem of incomplete predicted saliency map than CRFs, we elaborately design an accuracy performance comparison of the saliency produced using models with adversarial training and CRFs (i.e. LANet and S-CRF) on the MSRA-B dataset. Since our designed LANet consists of saliency predictor and adversarial network, to fairly evaluate the performance of CRFs, we design another comparison method, named S-CRF, which consists of saliency predictor for generating saliency map and a fully connected CRF for saliency refinement. As a result, S-CRF yields maximum F-measure $= 0.918$ and MAE $= 0.031$, while maximum F-measure and MAE of our LANet are 0.929 and 0.023, respectively. Overall, our LANet with adversarial training achieves the better performance than S-CRF with CRF in all the metrics. Therefore, adversarial training is more effective than CRFs for alleviating this problem.

Effectiveness of multi-scale contrast module. To better show the strength of our proposed multi-scale contrast module, we provide performance comparisons in terms of accuracy and efficiency, which are also illustrated in Fig. 8 and Table 3. Compared with the models without multi-scale contrast module, both MSC and LANet produce more pronounced gains in accuracy observed on all the six datasets, which demonstrates the capability of multi-scale contrast module to discovery and understand subtle visual contrast among multi-scale feature maps.

Furthermore, instead of the point-wise convolutions 1×1, our multi-scale contrast module utilizes depth-wise separable convolutions to reduce the dimensions of the features while filtering out the most important information, and yields about 1.06% (on average) improvement in F-measure. On the other hand, Table 3 lists out the efficiency comparison between the models with and without multi-scale contrast module. As can be seen, it introduces fairly low additional overhead: very small number of extra parameters and a negligible extra computation cost,and in return brings in remarkable additional performance gain.

5. Conclusion

In this paper, we have presented a lightweight adversarial network for salient object detection. Our proposed model introduces lightweight bottleneck blocks to significantly lower the computational cost and accelerate the process of training and inference. To enforce long-range spatial saliency contiguity, adversarial training is incorporated into saliency predictor, while consuming no additional computation cost at the test phase. Besides, to alleviate the limitation of contrast learning in contemporary CNN, we develop a multi-scale contrast module to rapidly and sufficiently capture local and global visual contrast. Comprehensive experiments demonstrate the superiority of our model in terms of accuracy and efficiency, and its brilliant potentials for real-time embedded applications.

Declaration of Competing Interest

We wish to draw the attention of the Editor to the following facts which may be considered as potential conflicts of interest and to significant financial contributions to this work.

We confirm that the manuscript has been read and approved by all named authors and that there are no other persons who satisfied the criteria for authorship but are not listed. We further confirm that the order of authors listed in the manuscript has been approved by all of us.

We confirm that we have given due consideration to the protection of intellectual property associated with this work and that there are no impediments to publication, including the timing of publication, with respect to intellectual property. In so doing we confirm that we have followed the regulations of our institutions concerning intellectual property.

We understand that the Corresponding Author is the sole contact for the Editorial process (including Editorial Manager and direct communications with the office). He/she is responsible for communicating with the other authors about progress, submissions of revisions and final approval of proofs. We confirm that we have provided a current, correct email address which is accessible by the Corresponding Author and which has been configured to accept email from huanglli3@mail2.sysu.edu.cn.

Acknowledgment

This work was supported in part by National Natural Science Foundation of China under Grant nos. 61906049 and 61702565, in part by the Science and Technology Program of Guangdong Province under Grant no. 2017B010116001, in part by the Fundamental Research Funds for the Central Universities under Grant no.18lgpy63, in part by Science and Technology Program of Guangzhou under Grant no. 201904010493 and in part by the Science and Technology Planning Project of Guangdong Province No. 2015B010128009.

References

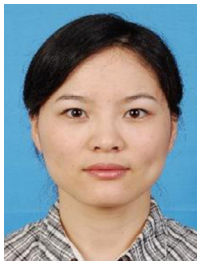
[1] A. Borji, S. Frintrop, D.N. Sihite, L. Itti, Adaptive object tracking by learning background context, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2012, IEEE, 2012, pp. 23–30.
[2] M.-M. Cheng, F.-L. Zhang, N.J. Mitra, X. Huang, S.-M. Hu, Repfinder: finding approximately repeated scene elements for image editing, in: Proceedings of the ACM Transactions on Graphics (TOG), 29, ACM, 2010, p. 83.
[3] G.-X. Zhang, M.-M. Cheng, S.-M. Hu, R.R. Martin, A shape-preserving approach to image resizing, in: Computer Graphics Forum, 28, Wiley Online Library, 2009, pp. 1897–1906.
[4] Y. Wei, X. Liang, Y. Chen, Z. Jie, Y. Xiao, Y. Zhao, S. Yan, Learning to segment with image-level annotations, Pattern Recognit. 59 (2016) 234–244.

[5] Y. Wei, J. Feng, X. Liang, M.-M. Cheng, Y. Zhao, S. Yan, Object region mining with adversarial erasing: a simple classification to semantic segmentation approach, in: Proceedings of the IEEE CVPR, 2017.
[6] Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.
[7] P. Krähenbühl, V. Koltun, Parameter learning and convergent inference for dense random fields, in: Proceedings of the International Conference on Machine Learning, 2013, pp. 513–521.
[8] G. Lin, C. Shen, A. Van Den Hengel, I. Reid, Efficient piecewise training of deep structured models for semantic segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 3194–3203.
[9] A.G. Schwing, R. Urtasun, Fully connected deep structured networks, arXiv:1503.02351 (2015).
[10] S. Zheng, S. Jayasumana, B. Romera-Paredes, V. Vineet, Z. Su, D. Du, C. Huang, P.H. Torr, Conditional random fields as recurrent neural networks, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1529–1537.
[11] A. Arnab, S. Jayasumana, S. Zheng, P.H. Torr, Higher order conditional random fields in deep neural networks, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 524–540.
[12] G. Li, Y. Yu, Visual saliency detection based on multiscale deep CNN features, in: IEEE Transactions on Image Processing, 25, 11th ed., IEEE, 2016, pp. 5012–5024.
[13] R. Zhao, W. Ouyang, H. Li, X. Wang, Saliency detection by multi-context deep learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 1265–1274.
[14] L. Wang, L. Wang, H. Lu, P. Zhang, X. Ruan, Saliency detection with recurrent fully convolutional networks, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 825–841.
[15] X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, J. Wang, Deepsaliency: multi-task deep neural network model for salient object detection, IEEE Trans. Image Process. 25 (8) (2016) 3919–3930.
[16] G. Li, Y. Yu, Contrast-oriented deep neural networks for salient object detection, in: IEEE Transactions on Neural Networks and Learning Systems, 29, 12th ed., IEEE, 2018, pp. 6038–6051.
[17] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, P. Torr, Deeply supervised salient object detection with short connections, in: Proceedings of the 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), IEEE, 2017, pp. 5300–5309.
[18] L. Zhang, J. Dai, H. Lu, Y. He, G. Wang, A bi-directional message passing model for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[19] H. Wang, L. Dai, Y. Cai, X. Sun, L. Chen, Salient object detection based on multi-scale contrast, Neural Networks 101 (2018) 47–56.
[20] J. Han, D. Zhang, C. Gong, N. Liu, X. Dong, Advanced deep-learning techniques for salient and category-specific object detection: a survey, IEEE Signal Process Mag. 35 (1) (2018) 84–100.
[21] D. Zhang, D. Meng, J. Han, Co-saliency detection via a self-paced multiple-instance learning framework, IEEE Trans. Pattern Anal. Mach. Intell. 39 (5) (2016) 865–878.
[22] D. Zhang, J. Han, L. Chao, J. Wang, X. Li, Detection of co-salient objects by looking deep and wide, Int. J. Comput. Vis. 120 (2) (2016) 215–232.
[23] J. Ba, R. Caruana, Do deep nets really need to be deep? in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 2654–2662.
[24] G. Hinton, O. Vinyals, J. Dean, Distilling the Knowledge in a Neural Network, Stat 1050 (2015) 9.
[25] A. Romero, N. Ballas, S.E. Kahou, A. Chassang, C. Gatta, Y. Bengio, Fitnets: hints for thin deep nets, arXiv:1412.6550 (2014).
[26] B. Hassibi, D.G. Stork, Second order derivatives for network pruning: Optimal brain surgeon, in: Proceedings of the Advances in Neural Information Processing Systems, 1993, pp. 164–171.
[27] Y. LeCun, J.S. Denker, S.A. Solla, Optimal brain damage, in: Proceedings of the Advances in Neural Information Processing Systems, 1990, pp. 598–605.
[28] S. Han, J. Pool, J. Tran, W. Dally, Learning both weights and connections for efficient neural network, in: Proceedings of the Advances in Neural Information Processing Systems, 2015, pp. 1135–1143.
[29] S. Han, J. Pool, S. Narang, H. Mao, S. Tang, E. Elsen, B. Catanzaro, J. Tran, W.J. Dally, DSD: regularizing deep neural networks with dense-sparse-dense training flow, arXiv:1607.04381 3(6) (2016).
[30] Y. Guo, A. Yao, Y. Chen, Dynamic network surgery for efficient DNNS, in: Proceedings of the Advances In Neural Information Processing Systems, 2016, pp. 1379–1387.
[31] H. Li, A. Kadav, I. Durdanovic, H. Samet, H.P. Graf, Pruning filters for efficient convnets, arXiv:1608.08710 (2016).
[32] K. Ahmed, L. Torresani, Connectivity learning in multi-branch networks, arXiv:1709.09582 (2017).
[33] T. Veniat, L. Denoyer, Learning time/memory-efficient deep architectures with budgeted super networks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 3492–3500.
[34] J. Bergstra, Y. Bengio, Random search for hyper-parameter optimization, J. Mach. Learn. Res. 13(Feb) (2012) 281–305.
[35] J. Snoek, O. Rippel, K. Swersky, R. Kiros, N. Satish, N. Sundaram, M. Patwary, M. Prabhat, R. Adams, Scalable bayesian optimization using deep neural networks, in: Proceedings of the International Conference on Machine Learning, 2015, pp. 2171–2180.
[36] W. Einhäuser, P. König, Does luminance-contrast contribute to a saliency map for overt visual attention? Eur. J. Neurosci. 17 (5) (2003) 1089–1097.
[37] D. Parkhurst, K. Law, E. Niebur, Modeling the role of salience in the allocation of overt visual attention, Vis. Res. 42 (1) (2002) 107–123.
[38] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Object detectors emerge in deep scene CNNS, arXiv:1412.6856 (2014).
[39] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial nets, in: Proceedings of the Advances in Neural Information Processing Systems, 2014, pp. 2672–2680.
[40] P. Luc, C. Couprie, S. Chintala, J. Verbeek, Semantic segmentation using adversarial networks, arXiv:1611.08408 (2016).
[41] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, L.-C. Chen, Mobilenetv2: inverted residuals and linear bottlenecks, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 4510–4520.
[42] A.G. Howard, M. Zhu, B. Chen, D. Kalenichenko, W. Wang, T. Weyand, M. Andreetto, H. Adam, Mobilenets: efficient convolutional neural networks for mobile vision applications, arXiv:1704.04861 (2017).
[43] M.-M. Cheng, N.J. Mitra, X. Huang, P.H. Torr, S.-M. Hu, Global contrast based salient region detection, IEEE Trans. Pattern Anal. Mach. Intell. 37 (3) (2015) 569–582.
[44] T. Liu, Z. Yuan, J. Sun, J. Wang, N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, IEEE Trans. Pattern Anal. Mach. Intell. 33 (2) (2011) 353–367.
[45] F. Perazzi, P. Krähenbühl, Y. Pritch, A. Hornung, Saliency filters: contrast based filtering for salient region detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, IEEE, 2012, pp. 733–740.
[46] L. Jiang, A. Koch, A. Zell, Salient regions detection for indoor robots using RGB-D data, in: Proceedings of the IEEE International Conference on Robotics and Automation (ICRA), 2015, IEEE, 2015, pp. 1323–1328.
[47] L. Wang, H. Lu, X. Ruan, M.-H. Yang, Deep networks for saliency detection via local estimation and global search, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2015, pp. 3183–3192.
[48] G. Lee, Y.-W. Tai, J. Kim, Deep saliency with encoded low level distance map and high level features, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 660–668.
[49] G. Li, Y. Xie, L. Lin, Y. Yu, Instance-level salient object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, IEEE, 2017, pp. 247–256.
[50] S. Xie, Z. Tu, Holistically-nested edge detection, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 1395–1403.
[51] T. Che, Y. Li, A.P. Jacob, Y. Bengio, W. Li, Mode regularized generative adversarial networks, arXiv:1612.02136 (2016).
[52] J. Zhao, M. Mathieu, Y. LeCun, Energy-based generative adversarial network, arXiv:1609.03126 (2016).
[53] C. Ledig, L. Theis, F. Huszár, J. Caballero, A. Cunningham, A. Acosta, A.P. Aitken, A. Tejani, J. Totz, Z. Wang, et al., Photo-realistic single image super-resolution using a generative adversarial network., in: Proceedings of the CVPR, 2, 2017, p. 4.
[54] J. Li, X. Liang, Y. Wei, T. Xu, J. Feng, S. Yan, Perceptual generative adversarial networks for small object detection, in: Proceedings of the IEEE CVPR, 2017.
[55] J. Pan, C.C. Ferrer, K. McGuinness, N.E. O'Connor, J. Torres, E. Sayrol, X. Giro-i Nieto, Salgan: visual saliency prediction with generative adversarial networks, arXiv:1701.01081 (2017).
[56] H. Pan, H. Jiang, Supervised adversarial networks for image saliency detection, arXiv:1704.07242 (2017).
[57] D. Zhu, L. Dai, Y. Luo, G. Zhang, X. Shao, L. Itti, J. Lu, Multi-scale adversarial feature learning for saliency detection, Symmetry (Basel) 10 (10) (2018) 457.
[58] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Proceedings of the Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
[59] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, arXiv:1409.1556 (2014).
[60] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, A. Rabinovich, Going deeper with convolutions, in: Proceedings of the IEEE Conference on Computer Vision and pattern recognition, 2015, pp. 1–9.
[61] K. He, X. Zhang, S. Ren, J. Sun, Deep residual learning for image recognition, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 770–778.
[62] X. Zhang, X. Zhou, M. Lin, J. Sun, Shufflenet: An extremely efficient convolutional neural network for mobile devices, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2018, pp. 6848–6856.
[63] S. Changpinyo, M. Sandler, A. Zhmoginov, The power of sparsity in convolutional neural networks, arXiv:1702.06257 (2017).
[64] M. Wang, B. Liu, H. Foroosh, Design of efficient convolutional layers using single intra-channel convolution, topological subdivisioning and spatial" bottleneck" structure, arXiv:1608.04337 (2016).
[65] F. CholletXception, Deep learning with depthwise separable convolutions, in: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, IEEE, 2017, pp. 1251–1258.
[66] H. Zhao, J. Shi, X. Qi, X. Wang, J. Jia, Pyramid scene parsing network, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2881–2890.
[67] K. He, X. Zhang, S. Ren, J. Sun, Spatial pyramid pooling in deep convolutional networks for visual recognition, in: Proceedings of the European Conference on Computer Vision, Springer, 2014, pp. 346–361.

[68] P.O. Pinheiro, T.-Y. Lin, R. Collobert, P. Dollár, Learning to refine object segments, in: Proceedings of the European Conference on Computer Vision, Springer, 2016, pp. 75–91.
[69] T. Liu, J. Sun, N.-N. Zheng, X. Tang, H.-Y. Shum, Learning to detect a salient object, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2007, CVPR'07, IEEE, 2007, pp. 1–8.
[70] J. Shi, Q. Yan, L. Xu, J. Jia, Hierarchical image saliency detection on extended CSSD, IEEE Trans. Pattern Anal. Mach. Intell. 38 (4) (2016) 717–729.
[71] C. Yang, L. Zhang, H. Lu, X. Ruan, M.-H. Yang, Saliency detection via graph-based manifold ranking, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 3166–3173.
[72] V. Movahedi, J.H. Elder, Design and perceptual validation of performance measures for salient object segmentation, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2010, IEEE, 2010, pp. 49–56.
[73] D. Martin, C. Fowlkes, D. Tal, J. Malik, A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics, in: Proceedings of the Eighth IEEE International Conference on Computer Vision, 2001, ICCV 2001., 2, IEEE, 2001, pp. 416–423.
[74] Y. Li, X. Hou, C. Koch, J.M. Rehg, A.L. Yuille, The secrets of salient object segmentation, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 280–287.
[75] P. Jiang, H. Ling, J. Yu, J. Peng, Salient region detection by UFO: uniqueness, focusness and objectness, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 1976–1983.
[76] R. Achanta, S. Hemami, F. Estrada, S. Susstrunk, Frequency-tuned salient region detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009, IEEE, 2009, pp. 1597–1604.
[77] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, et al., Tensorflow: a system for large-scale machine learning., in: Proceedings of the OSDI, 16, 2016, pp. 265–283.
[78] X. Chen, A. Zheng, L. Jia, L. Feng, Look, perceive and segment: finding the salient objects in images via two-stream fixation-semantic CNNS, in: Proceedings of the IEEE International Conference on Computer Vision, 2017.
[79] T. Wang, A. Borji, L. Zhang, P. Zhang, H. Lu, A stagewise refinement model for detecting salient objects in images, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
[80] Z. Luo, A. Mishra, A. Achkar, J. Eichel, P.M. Jodoin, Non-local deep features for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017.
[81] P. Zhang, D. Wang, H. Lu, H. Wang, X. Ruan, Amulet: aggregating multi-level convolutional features for salient object detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
[82] P. Zhang, D. Wang, H. Lu, H. Wang, B. Yin, Learning uncertain convolutional features for accurate saliency detection, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017.
[83] X. Zhang, T. Wang, J. Qi, H. Lu, G. Wang, Progressive attention guided recurrent network for salient object detection, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.
[84] X. Li, F. Yang, H. Cheng, W. Liu, D. Shen, Contour knowledge transfer for salient object detection, in: Proceedings of the ECCV, 2018.
[85] Q. Yao, M. Feng, H. Lu, G.W. Cottrell, Hierarchical cellular automata for visual saliency, Int. J. Comput. Vis. 126 (7) (2018) 1–20.
[86] S. Chen, X. Tan, B. Wang, X. Hu, Reverse attention for salient object detection, in: Proceedings of the European Conference on Computer Vision, 2018.

Lili Huang is a Ph.D. student in Computer Science with the School of Data and Computer Science, Sun Yat-sen University, China. Before that, she received her B.E. and master degrees from School of Computer Science and Technology, Anhui University, China. Her current research interests lie in the fields of computer vision, machine learning, and cognitive science. She was the recipient of the Best Paper Award in IEEE ICME 2017.

Guanbin Li is currently a research associate professor in School of Data and Computer Science, Sun Yat-sen University. He received his Ph.D. degree from the University of Hong Kong, China in 2016. He was a recipient of Hong Kong Postgraduate Fellowship. His current research interests include computer vision, image processing, and deep learning. He has authorized and co-authorized on more than 20 papers in top-tier academic journals and conferences. He serves as an area chair for the conference of VISAPP. He has been serving as a reviewer for numerous academic journals and conferences such as TPAMI, TIP, TMM, TC, CVPR2018 and IJCAI2018.



Ya Li is a lecturer in School of Computer Science and Educational Software, Guangzhou University, Guangzhou, China. She received the B.E. degree from Zhengzhou University, Zhengzhou, China, in 2002, M.E. degree from Southwest Jiaotong University, Chengdu, China, in 2006 and Ph.D. degree from Sun Yat-sen University, Guangzhou, in 2015. Her current research focuses on computer vision and machine learning.

Liang Lin is a full Professor of Sun Yat-sen University. He is the Excellent Young Scientist of the National Natural Science Foundation of China. From 2008 to 2010, he was a Post-Doctoral Fellow at the University of California, Los Angeles. From 2014 to 2015, as a senior visiting scholar, he was with the Hong Kong Polytechnic University and the Chinese University of Hong Kong. He has authored and co-authored more than 100 papers in top-tier academic journals and conferences. He has been serving as an associate editor of IEEE Trans. Human-Machine Systems, The Visual Computer and Neurocomputing. He served as Area/Session Chair for numerous conferences, including ICME, ACCV, and ICMR. He was the recipient of the Best Paper Runners-Up Award at ACM NPAR 2010, a Google Faculty Award in 2012, the Best Paper Diamond Award at IEEE ICME 2017, and the Hong Kong Scholars Award in 2014. He is a Fellow of IET.