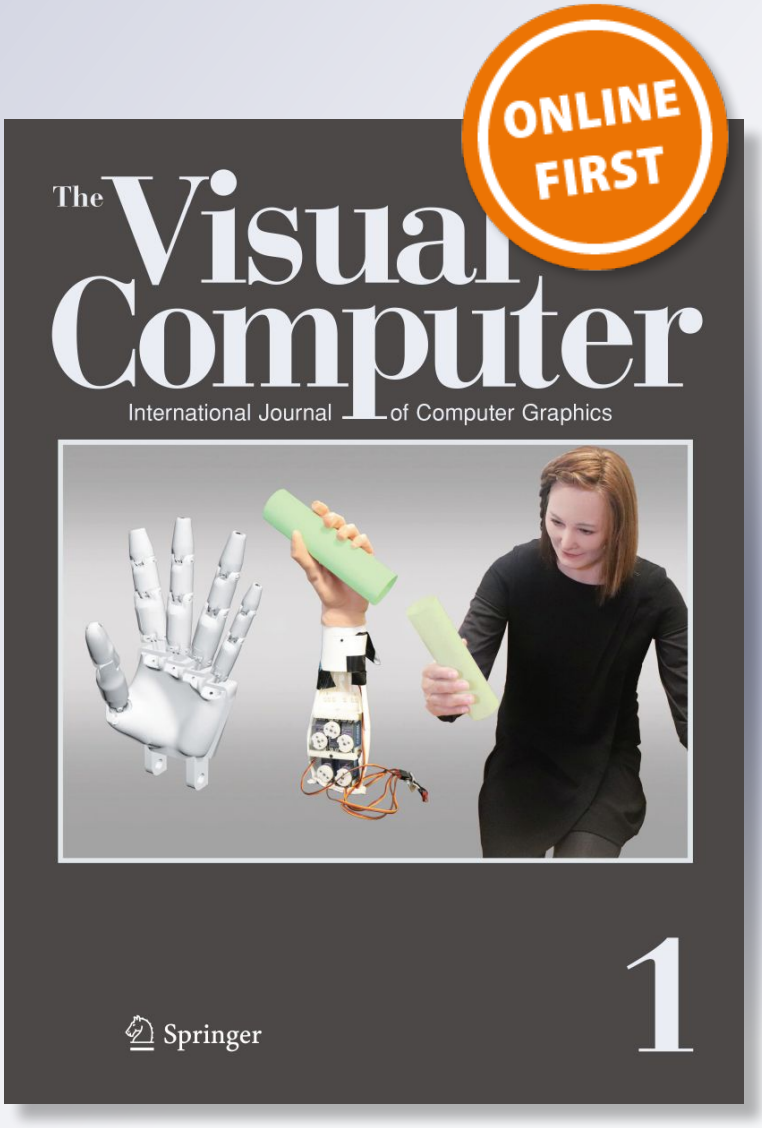
Salient object detection via hybrid upsampling and hybrid loss computing

Zhengyi Liu, Jiting Tang & Peng Zhao

The Visual Computer
International Journal of Computer Graphics

ISSN 0178-2789

Vis Comput
DOI 10.1007/s00371-019-01659-w



Springer

The Visual Computer
https://doi.org/10.1007/s00371-019-01659-w

ORIGINAL ARTICLE

Salient object detection via hybrid upsampling and hybrid loss computing

Zhengyi Liu[1,2] · Jiting Tang[1,2] · Peng Zhao[1,2]

Abstract
Salient object detection aims to detect distinct objects which attract human most. It has achieved substantial progress using deep convolutional neural networks in which conventional deconvolution operation is used as recovering the size of image in dense prediction tasks and cross-entropy loss is applied to compute the difference between saliency map and ground truth in pixel level. Different from conventional deconvolution operation, hybrid upsampling block is proposed to retain the detail of object by increasing the receptive field and spatial information when recovering the image size, and hybrid loss which consists of cross-entropy loss and area loss is proposed to train the network optimized by area constraint. At last, an encoder-decoder network based on hybrid upsampling block and hybrid loss is implemented in public benchmark dataset and achieves the best performance against state-of-the-art methods.

Keywords Salient object detection · Hybrid upsampling · Hybrid loss · CRF · Area constraint

1 Introduction

Humans can perceive their visual environment at a glance attribute to important visual attention mechanism. It can guide the observer to gaze salient and informative locations in the visual field or identify and locate distinctive objects which attract human most. They are called eye-fixation prediction [20,21,27] and salient object detection (SOC) [28,34,38], respectively, and regarded as a prerequisite step to narrow down subsequent more complex vision tasks especially semantic segmentation task [30,31]. Salient object detection is probed in the paper. The initial research focuses on SOD in an image, then it is extended to RGBD saliency detection [4,5] in which depth cue can be utilized , co-saliency detection [8,37] in which inner- and intersaliency constraint need be considered simultaneously, and video saliency detection [11,23,29,30,33] in which temporal and spatial relationship is explored. Our work focuses on salient object detection in an image.

In the last 2 decades, many SOD methods from low-level heuristic view have emerged, but with the development of Convolutional Neural Networks (CNNs), SOD methods based on deep learning have achieved more excellent performance. In most of network structures, they adopt conventional convolution operation for recovering image size. Zhang et al. [39] proposed a hybrid upsampling block, which consists of conventional deconvolution and linear interpolation with 1×1 convolution. Inspired by it, we designed different hybrid upsampling block which ensures more precision when recovering image size by deconvolution with bilinear interpolation kernel, atrous convolution and concatenation operation. On the other hand, the training of network is related to loss function. Cross-entropy loss which means the error between saliency map and ground truth is the most popular loss function. Luo et al. [19] proposed a boundary loss function to reduce errors on the boundary. Different from it, hybrid loss which is composed of area loss and cross-entropy loss is designed to improve network performance by pixel-level and area-level constraint. Based on proposed hybrid upsampling block and hybrid loss, salient object detection network with encoder-decoder fully convolutional network

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

Jiting Tang
1796340141@qq.com

Peng Zhao
18868519@qq.com

1 Key Laboratory of Intelligent Computing Signal Processing of Ministry of Education, Anhui University, Hefei, China

2 School of Computer Science and Technology, Anhui University, Hefei, China

Published online: 09 May 2019

framework is designed and implemented. Our encoder part is completely consistent with the convolution part of VGG-16 net and discards the last fully connected layers. Decoder part employs hybrid upsampling operation instead of common deconvolution operation to recover the size of image, increase the receptive field and spatial information. Our model has several unique features, as outlined below:

- Hybrid upsampling block is proposed for replacing conventional deconvolutional operation to recover the size of image in dense salient object detection task. It consists of deconvolution with bilinear interpolation kernel which increases the size of the input feature map, atrous convolution with different rates which increases the receptive field and spatial information and concatenation operation which can refine feature outline by combining original feature and feature after upsampling.
- Hybrid loss function which consists of cross-entropy loss function and area loss function is proposed to train our network. In this way, the effect of network training is better, the real and predicted values are closer. Experimental results show that hybrid loss function can achieve better training effects by reducing the errors between the saliency map and ground truth than cross-entropy loss.
- An encoder-decoder fully convolutional neural network based on hybrid upsampling and hybrid loss function is proposed to detect salient object in an image. Experiments demonstrate it outperforms the state-of-the-art methods.

2 Related work

Salient object detection aims to highlight entire objects of interest and suppress background regions. It can be divided into down-top methods and top-down methods. Down-top methods utilize low-level handcrafted feature and prior knowledge to measure saliency. For example, Wei et al. [32] exploited two common priors about backgrounds in natural images to provide more clues for the salient object problem. Lu et al. [36] proposed manifold ranking techniques to detect salient objects. Zhu et al. [40] proposed a robust background measure to achieve saliency optimization. Wang et al. [26] jointly ranking learning and subspace learning to find salient objects.

But down-top methods have been outperformed by deep learning approaches in recent years. With CNNs, saliency detection problem has been redefined as a labeling prediction problem where feature is selected automatically through gradient descent. Li et al. [15] proposed a multiple scale network architecture for salient object detection. Lu et al. [25] proposed a saliency detection algorithm by integrating both local estimation and global search. Li et al. [17] proposed a multitask deep neural network model for salient object detection. Shen et al. [28] proposed a saliency transfer method by taking the advantage of the existing large annotated datasets for identifying the primary and smooth connected salient object areas in an image. Lee et al. [14] utilized both high-level and low-level features for saliency detection under a unified deep learning framework. Li et al. [16] proposed a pixel-level fully convolutional stream and a segment-wise spatial pooling stream for salient object detection. Zhang et al. [39] proposed a deep fully convolutional network model for accurate salient object detection by learning deep uncertain convolutional features and utilizing hybrid upsampling method.

Conventional deconvolution operation is adopted in aforementioned networks to recover image size for dense salient object detection task. However, Lu et al. [39] proposed a hybrid upsampling method which consists of conventional deconvolution operation and linear interpolation. Different from it, we designed a hybrid upsampling block which is composed of deconvolution with bilinear interpolation kernel, atrous convolution and concatenation operation in decoder part to retain more details when recovering the size of feature. As we all know, the training of network is related to loss function. Cross-entropy computes the error between saliency map and ground truth. Luo et al. [19] proposed a boundary loss function to reduce errors on the boundary. Different from it, hybrid loss consists of area loss and cross-entropy loss function is proposed to improve network performance and reduce the difference between saliency map and ground truth in pixel-level and area-level simultaneously.

3 The proposed method

3.1 Overview of network architecture

Our network is composed of an encoder fully convolutional network (FCN) for deep feature extraction, and a corresponding decoder FCN for low-level information reconstruction, as illustrated in Fig. 1. Hybrid upsampling block consists of deconvolution with bilinear interpolation kernel, atrous convolution operation with different rates and concatenation operation. It is used to recover the size of the feature map, increase the receptive field and spatial information. Meanwhile, hybrid loss function is computed between the saliency map and the ground truth at the pixel-level and area-level.

The algorithm for network model is shown in Algorithm 1. The input image is first reshaped into 352×352 size. It is then fed into encoder part which adopts CONV-1 to CONV-5 derived from VGG-16 Network [22], which discards the last two fully connected layers for dense image saliency detection task. The convolution blocks with kernel size $k = 3$ and stride $= 1$ in the encoder part are used to extract the fea-

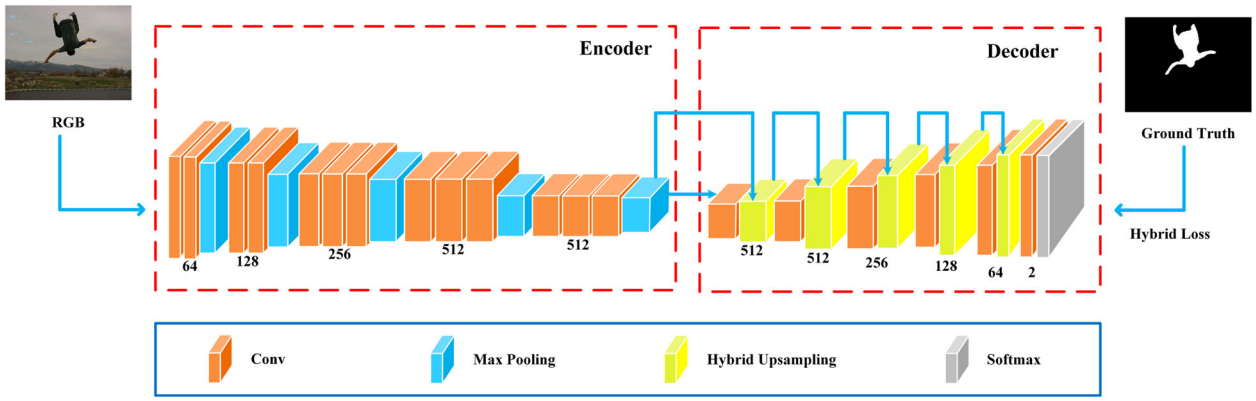


Fig. 1 Salient object detection network framework based on hybrid upsampling and hybrid loss

tures of input image. The pooling layers with kernel size $k = 2$ and stride = 2 in the encoder part are used to reduce the dimensionality of feature map, as shown in Table 1. Decoder part is composed of five deconvolutional blocks (DECONV-1 to DECONV-5) for not only recovering the size of the feature map but also retaining details of salient object, and convolution operation and a softmax function for generating the saliency map. Each deconvolutional block is convolutional filters with kernel size $k = 1$ followed by a hybrid upsampling block. It increases nonlinear mapping for fitting the network better. Hybrid upsampling which is inspired by residual network [24] uses the output from CONV-5 or previous deconvolution block as input to generate output by deconvolution with bilinear interpolation kernel, atrous convolution [6] and concatenation operations.

Algorithm 1: The Algorithm for Our Network Model

Input: I is a RGB image with 352*352 size, Y is a ground truth image with 352*352 size;
Output: $\hat{I}$ is a corresponding saliency map with 352*352 size;

$loss = +\infty$;
while $loss > 0$ **do**
  $F(0) = I$;
  $ConvLevel = 5$;
  for $i = 1$ *to* 2 **do**
    $F(i) = Maxpooling(Conv(Conv(F(i-1))))$;
  for $i = 3$ *to* $ConvLevel$ **do**
    $F(i) = Maxpooling(Conv(Conv(Conv(F(i-1)))))$;
  $Q(0) = F(5)$;
  $DeConvLevel = 5$;
  for $j = 1$ *to* $DeConvLevel$ **do**
    $B_{11} = Conv(Q(j-1))$;
    $B_{12} = DeconvBilinearInterpolation(B_{11})$;
    $B_{13} = AtrousConv(B_{12}, 2)$;
    $B_{14} = AtrousConv(B_{12}, 6)$;
    $B_{15} = Conv(Resize(Q(j-1)))$;
    $Q(j) = Concatenate(B_{12}, B_{13}, B_{14}, B_{15})$;
  $\hat{I} = Softmax(Conv(Q(5)))$;
  $loss = HybridLoss(\hat{I}, Y)$;

3.2 Hybrid upsampling block

The resolution of feature after convolutional operation in five blocks becomes very small, so recovering the size of feature and retaining object details are the task of decoder in salient object detection. As the basic and important unit, hybrid upsampling block needs to conduct deconvolution operation to recover the size of feature, but feature maybe miss some information due to padding operation, so original feature after resizing and convolutional operation can be added into deconvolutional feature as supplementary information. Meanwhile, atrous convolution can enlarge the receptive field of feature and increase the spatial information. The feature after atrous convolution can also provide additional information.

Therefore, hybrid upsampling block is served as refining the detail of the salient object in the process of recovering the size of image. Figure 2 illustrates the detail about hybrid upsampling block. The input is first performed by the convolution with the kernel size k = 1 to form feature B_{11}, and it is conducted by deconvolution operation with bilinear interpolation kernel to generate B_{12} which increases the size of the input feature map. And then, two atrous convolution operations with rate = (2, 6), respectively, are carried on B_{12} to generate B_{13}, B_{14} to enlarge the receptive field of the feature. And then, the input feature is resized and then processed by the convolution to generate B_{15}. Four parts B_{12}, B_{13}, B_{14}, B_{15} with the same resolution are last concatenated together as output. The output of previous deconvolution block is used as input for the next deconvolution block.

When the network uses deconvolution operation to recover the image size, the deconvolutional kernel plays an important role in the task of image recovery. Deconvolutional kernel with bilinear interpolation is adopted in our work. It

Table 1 This is parameters

Block	Layer	K	P	S	Output
CONV-1	2conv	3×3	Yes	1	$352 \times 352 \times 64$
	max-pool	2×2	Yes	2	$176 \times 176 \times 64$
CONV-2	2conv	3×3	Yes	1	$176 \times 176 \times 128$
	max-pool	2×2	Yes	2	$88 \times 88 \times 128$
CONV-3	3conv	3×3	Yes	1	$88 \times 88 \times 256$
	max-pool	2×2	Yes	2	$44 \times 44 \times 256$
CONV-4	3conv	3×3	Yes	1	$44 \times 44 \times 512$
	max-pool	2×2	Yes	2	$22 \times 22 \times 512$
CONV-5	3conv	3×3	Yes	1	$22 \times 22 \times 512$
	max-pool	2×2	Yes	2	$11 \times 11 \times 512$
DECONV-1	conv	1×1	Yes	1	$11 \times 11 \times 128$
	hybrid-upsample	Multikernel	Yes	1	$22 \times 22 \times 512$
DECONV-2	conv	1×1	Yes	1	$22 \times 22 \times 128$
	hybrid-upsample	Multikernel	Yes	1	$44 \times 44 \times 512$
DECONV-3	conv	1×1	Yes	1	$44 \times 44 \times 64$
	hybrid-upsample	Multikernel	Yes	1	$88 \times 88 \times 256$
DECONV-4	conv	1×1	Yes	1	$88 \times 88 \times 32$
	hybrid-upsample	Multikernel	Yes	1	$176 \times 176 \times 128$
DECONV-5	conv	1×1	Yes	1	$176 \times 176 \times 16$
	hybrid-upsample	Multikernel	Yes	1	$352 \times 352 \times 64$
CONV	conv + softmax	1×1	No	1	$352 \times 352 \times 2$

Details the proposed deep convolutional network for detecting salient objects (K kernel, P padding, S strides)

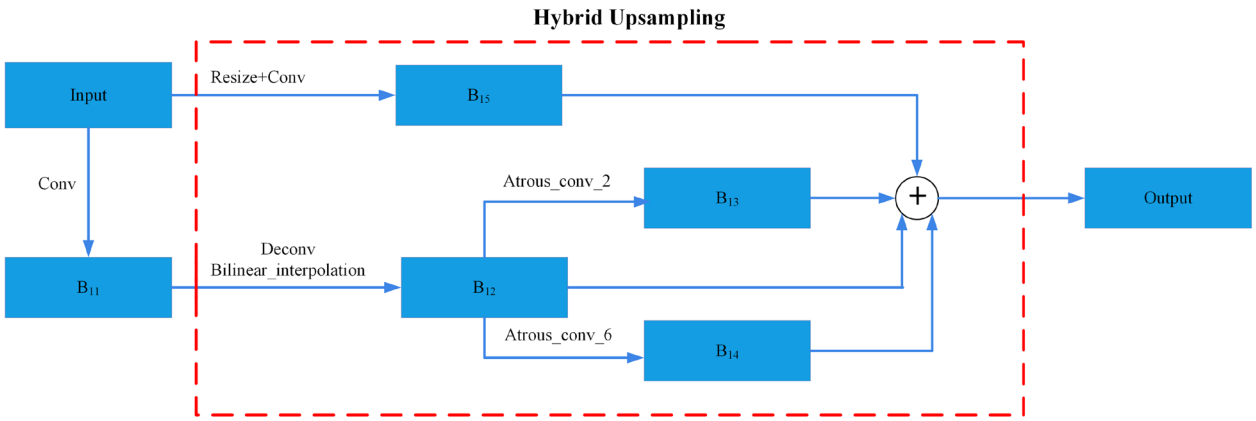


Fig. 2 Details of hybrid upsampling block

is very similar to convolution with bilinear interpolation kernel [39]. The feature map reconstructed with deconvolution with bilinear interpolation kernel can recover the features of the original feature smoother. Bilinear interpolation is the linear interpolation from X and Y directions simultaneously. It is done first on the X-axis direction and then on the Y-axis direction. From a mathematical point of view, suppose $Q_{11} = (x_1, y_1)$, $Q_{12} = (x_1, y_2)$, $Q_{21} = (x_2, y_1)$ and $Q_{22} = (x_2, y_2)$, the value of function f at the point $P(x, y)$ is computed first by linear interpolation in x direction to form $f(x, y_1)$, $f(x, y_2)$:

$$f(x, y_1) \approx \frac{x_2 - x}{x_2 - x_1} \cdot f(Q_{11}) + \frac{x - x_1}{x_2 - x_1} \cdot f(Q_{21}) \quad (1)$$

$$f(x, y_2) \approx \frac{x_2 - x}{x_2 - x_1} \cdot f(Q_{12}) + \frac{x - x_1}{x_2 - x_1} \cdot f(Q_{22}) \quad (2)$$

where $f(Q_{11})$, $f(Q_{12})$, $f(Q_{21})$, $f(Q_{22})$ denote the value of function f at the point Q_{11}, Q_{12}, Q_{21}, Q_{22}, and then the

value of function f at the point $P(x, y)$ is get by linear interpolation in y direction:

$$f(P) \approx \frac{y_2 - y}{y_2 - y_1} \cdot f(x, y_1) + \frac{y - y_1}{y_2 - y_1} \cdot f(x, y_2). \tag{3}$$

By deconvolution operation with the linear interpolation from two direction, deconvolution feature is smoother, and further fused with atrous convolutional feature for larger receptive field, and further complemented by original feature for recovering more details of the feature.

3.3 Hybrid loss function

Hybrid loss function is composed of pixel-level cross-entropy function and area-level area loss function. Cross-entropy loss is used to determine how close the predicted value is to the true value. In other words, the smaller cross-entropy loss is, the closer the predicted value and the true value are. Cross-entropy is often used to combine with softmax function, which is introduced by Tensorflow framework [1], to generate saliency value. It is defined as:

$$H(y(I), \hat{y}(I)) = -\frac{1}{N} \cdot \sum_{i=1}^{N} [y_i log(\hat{y}_i(I)) + (1 - y_i) log(1 - \hat{y}_i(I))] \tag{4}$$

where $I = \{I_i | i = 1, \ldots, N\}$ is an input image, N is the number of pixels of input image I, $y(I)$ is ground truth, $\hat{y}(I)$ is predicted saliency map.

Inspired by the boundary loss function proposed by [19], area loss function is proposed. Area loss is used to determine how close predicted salient region is to the real salient region. In other words, the closer predicted salient region and the true salient region are, the larger their intersection is, and the smaller area loss is. It is defined as:

$$\text{Area}(y(I), \hat{y}(I)) = 1 - \alpha * \frac{\sum_{i=1}^{N} (y(I_i) \bigcap \hat{y}(I_i)) + \delta}{\sum_{i=1}^{N} y(I_i) + \sum_{i=1}^{N} \hat{y}(I_i) + \gamma} \tag{5}$$

where $\bigcap$ denotes point-wise multiplication operator, and the constant $\alpha = 2$, $\delta = 1$,$\gamma = 1$ are set for easing the gradient-based training process and resulting in a faster convergence. Compared with boundary loss [19], area loss needs no training process for computing the boundary of predicted salient objects, it only sums all the saliency value in the ground truth ($y(I)$), all the saliency value in predicted saliency map ($\hat{y}(I)$) and all the saliency value in their point-wise multiplication result $y(I) \bigcap \hat{y}(I)$. Compared with boundary loss [19], area loss is also easy to convergent by adding constants in the molecular and denominator of Eq. 5. It will not give arise to gradient explosion problem while boundary loss can.

Area loss function is complementary to cross-entropy loss function for the fewer errors between the saliency map and the ground truth. Therefore, hybrid loss function can be computed as:

$$\text{Loss}(y(I), \hat{y}(I)) = \text{Area}(y(I), \hat{y}(I)) + H(y(I), \hat{y}(I)) \tag{6}$$

The whole loss computation procedure is end-to-end trainable. In this way, the effect of network training is better, and the ground truth and predicted values are closer.

3.4 Refinement of salient object detection

Due to the lack of edge information optimization in the saliency map obtained from deep learning, we use Conditional Random Fields (CRF) model [13] to optimize saliency map and improve spatial coherence. It is a conditional probability distribution model. Image pixel labels are optimized with respect to the following energy function of the CRF:

$$E(y) = -\sum_{i} \log P(y_i) + \sum_{i,j} \theta_{ij}(y_i, y_j) \tag{7}$$

where y represents a complete label assignment for all pixels and $P(y_i)$ is the probability of pixel x_i being assigned with the label prescribed by y_i, $\theta_{ij}(y_i, y_j)$ is a pair-wise potential and defined as:

$$\theta_{ij} = \mu(y_i, y_j) \left[\omega_1 exp\left(-\frac{\|p_i - p_j\|^2}{2\delta_\alpha^2} - \frac{\|I_i - I_j\|^2}{2\delta_\beta^2} \right) + \omega_2 exp\left(-\frac{\|p_i - p_j\|^2}{2\delta_\gamma^2} \right) \right] \tag{8}$$

where $\mu(y_i, y_j) = 1$ if $y_i \neq y_j$, and zero otherwise. θ_{ij} involves two kernels. The first kernel depends on pixel positions (p) and pixel intensities (I) and encourages nearby pixels with similar colors to take similar salient instance labels, while the second kernel only considers spatial proximity when enforcing smoothness. The hyperparameters δ_α, δ_β and δ_γ control the scale of Gaussian kernels. The parameters ω_1, ω_2, δ_α, δ_β and δ_γ are set to 4.0, 3.0, 49.0, 5.0 and 3.0, respectively, in our experiments.

4 Experimental results

4.1 Experimental setting

We evaluate the performance of our method on four public datasets.

MSRA-B [18] MSRA-B is a subset of MSRA10k [7]. It only has 5000 images, and they are all relatively simple images. Most of the images have a single salient object.

DUT-OMRON [36] DUT-OMRON contains 5168 challenging images, each of which has one or more salient objects and relatively complex backgrounds.

HKU-IS [15] HKU-IS is the large dataset containing 4447 challenging images, most of which have either low contrast or multiple salient objects.

ECSSD [35] ECSSD contains 1000 images with complex structure acquired from the Internet.

In this paper, we divide the MSRA-B dataset into three parts as described in [12], 2500 for training, 500 for validation and the remaining 2000 images for testing. HKU-IS dataset are also split into 2500 training images, 500 validation images and the remaining 1447 test images. The rest dataset are all used for testing. All the datasets contain manually annotated ground truth saliency maps. The training and validation set are augmented with horizontal flipping to train our model.

Our model is implemented in TensorFlow [1]. The weights in the CONV-1 to CONV-5 blocks are initialized with the pretrained weights of VGG-16 [22] net. All the weights of newly added convolution and deconvolution layers are initialized randomly with specified value $\delta = 0.01$, and the biases are initialized to 0.

4.2 Evaluation metrics

Evaluation metrics are used to evaluate the performance of different saliency models.

PR Curve For a saliency map, we first binarize it using a threshold to obtain the corresponding binary mark (B), and then compare the binary mask (B) with the corresponding ground truth (Y); finally, calculate the average precision and recall value in the whole dataset. The formula is as follows:

$$\text{Precision} = \frac{|Y \cap B|}{|B|}, \text{Recall} = \frac{|Y \cap B|}{|Y|} \tag{9}$$

F-measure F-measure is computed by:

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

where we set $\beta^2 = 0.3$ the same as [2]. The bigger F_β, the better result.

MAE [3] Mean absolute error (MAE) refers to the average pixel-wise error between the saliency map and ground truth. It is computed by:

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

where H and W are height and width of the image, $S(x, y)$ and $G(x, y)$ denote the saliency map and ground truth at the pixel (x, y), respectively. The smaller MAE, the better result.

S-measure [9] Structural similarity measure simultaneously evaluates region-aware and object-aware structural similarity between saliency map and ground truth. S-measure is defined as:

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

where S_o and S_r are the object-aware and region-aware structural similarity, respectively, λ is the balance parameter and set as 0.5 in our experiment. The bigger S, the better result.

In order to compare our model with the state-of-the-art models fairly, all the evaluation metrics use the same codes[1] provided by [10].

4.3 Ablation study

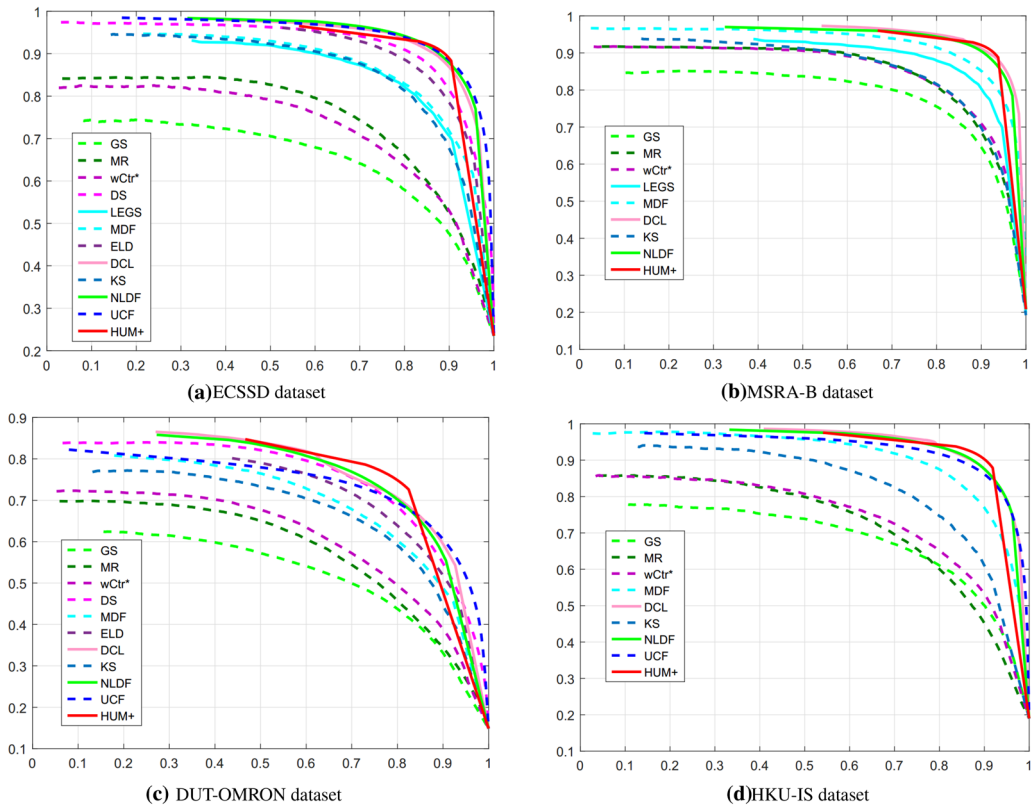
We verify the effectiveness of each contribution in our model. At first, baseline model is constructed from an encoder-decoder framework with conventional deconvolution operation and common cross-entropy loss function and denoted as baseline. Second, deconvolution operation is replaced with hybrid upsampling block, and it is denoted as HUM$^-$. Third, cross-entropy loss function is replaced with hybrid loss function, and it is denoted as HUM. Last, CRF optimization is added to the model HUM, and it is denoted as HUM$^+$.

From Table 2, we can see that model HUM$^-$ in the second line is superior to baseline model *Baseline* in the first line except for one evaluation metric F_β in HKU-IS dataset. The results verify the effectiveness of the first contribution about hybrid upsampling block. Meanwhile, we can see that the performance of model HUM in the third line is better than the model HUM$^-$ in the second line on all the evaluation metrics. Therefore, the results verify the effectiveness of the second contribution about hybrid loss. At last, we can see that the model HUM$^+$ in the fourth line outperforms the model HUM except for S-measure on ECSSD, MSRA-B and DUT-OMRON dataset. That is to say, CRF plays a limited role in improving the performance. But most evaluation metrics still have the impact, so it is adopted for optimization in our work.

4.4 Comparison with state-of-the-art methods

We compare our model against state-of-the-art methods, including Geodesic Saliency (GS) [32], Manifold Ranking (MR) [36], optimized Weighted Contrast (wCtr*) [40], Multi-task Deep Neural Network model for salient (DS) [17], Local Estimation and Global Search (LEGS) [25], Multiscale Deep Features (MDF) [15], Encoded Low-Level Distance Map (ELD) [14], Deep Contrast Learning (DCL) [16], Kernelized Subspace (KS) [26], Non-Local Deep Features (NLDF) [19] and uncertain Convolutional Features (UCF) [39]. We

[1] https://mmcheng.net/zh/code-data.



(a)ECSSD dataset **(b)**MSRA-B dataset

(c) DUT-OMRON dataset **(d)**HKU-IS dataset

Fig. 3 PR curves comparison of our method and other state-of-the-art methods

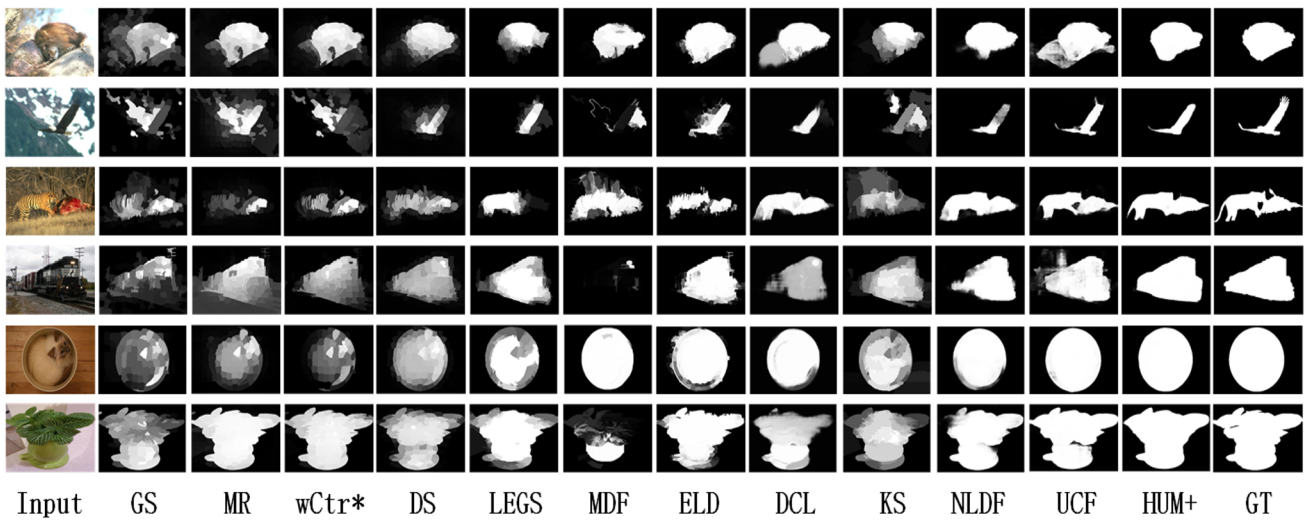


Fig. 4 Visual comparison of different model

use the result saliency maps provided by the authors for fairly comparison.

Table 3 and Fig. 3 demonstrate our method outperforms all existing salient object detection methods across the aforementioned public datasets in terms of F_β, MAE and S-measure evaluation metrics. Although there are intersections in Fig. 3 PR curve, Table 3 evaluation metrics clearly show the performance of our saliency map is superior to the others. Figure 4 provides a visual comparison of our method and other methods. It can be seen that our method generates more accurate saliency maps in various challenging cases. From visual comparison results, we find that the results from NLDF, UCF and ours are obviously superior to the others. In

Table 2 Ablation study about the role of hybrid upsampling, hybrid loss and CRF on baseline model

Models	ECSSD			MSRA-B			DUT-OMRON			HKU-IS		
	$F_\beta \uparrow$	MAE↓	S-measure↑	$F_\beta \uparrow$	MAE↓	S-measure↑	$F_\beta \uparrow$	MAE↓	S-measure↑	$F_\beta \uparrow$	MAE↓	S-measure↑
Baseline	0.8663	0.0758	0.8482	0.8757	0.0574	0.8697	0.7079	0.0859	0.7478	0.8731	0.0488	0.8730
HUM−:Baseline + hybrid upsampling	0.8716	0.0684	0.8709	0.8824	0.0496	0.8945	0.7284	0.0755	0.7931	0.8707	0.0467	0.8864
HUM:Baseline + hybrid upsampling + hybrid loss	0.8877	0.0588	**0.8787**	0.9000	0.0417	**0.9025**	0.7450	0.0704	**0.7958**	0.8874	0.0418	0.8899
HUM+:Baseline + hybrid upsampling + hybrid loss + CRF	**0.9000**	**0.0543**	0.8782	**0.9104**	**0.0379**	0.9022	**0.7633**	**0.0665**	0.7948	**0.9050**	**0.0369**	**0.8915**

The results of the top are shown in bold

Table 3 Quantitative comparison of different methods

Methods	ECSSD			MSRA-B			DUT-OMRON			HKU-IS		
	$F_\beta \uparrow$	MAE↓	S-measure↑	$F_\beta \uparrow$	MAE↓	S-measure↑	$F_\beta \uparrow$	MAE↓	S-measure↑	$F_\beta \uparrow$	MAE↓	S-measure↑
GS	0.6187	0.2058	0.6603	0.7179	0.1435	0.7598	0.5243	0.1732	0.6384	0.6391	0.1666	0.6937
MR	0.6962	0.1861	0.6924	0.7801	0.1265	0.7795	0.5764	0.1870	0.6456	0.6774	0.1715	0.6750
wCtr*	0.6777	0.1713	0.6886	0.7832	0.1103	0.7944	0.5944	0.1438	0.6815	0.6902	0.1406	0.7078
DS	0.8140	0.1216	0.8207	–	–	–	0.6839	0.1204	0.7502	–	–	–
LEGS	0.8016	0.118	0.7866	0.8314	0.0825	0.8412	–	–	–	–	–	–
MDF	0.8196	0.1049	0.7762	0.8231	0.1040	0.8565	0.6866	0.0916	0.7208	0.7804	0.1292	0.8181
ELD	0.8408	0.0783	0.8413	–	–	–	0.6699	0.0909	0.7514	–	–	–
DCL	0.8853	0.0678	0.8686	0.9004	0.0467	0.9014	0.7387	0.0797	0.7710	0.8873	0.0481	0.8770
KS	0.7796	0.1313	0.7636	0.7754	0.1151	0.7808	0.6444	0.1306	0.7075	0.7521	0.1193	0.7292
NLDF	0.8949	0.0625	0.8748	0.8993	0.0478	0.8943	0.7418	0.0796	0.7704	0.8897	0.0480	0.8782
UCF	0.8670	0.0689	**0.8836**	–	–	–	0.6760	0.1204	0.7599	0.8436	0.0612	0.8742
HUM+	**0.9000**	**0.0543**	0.8782	**0.9104**	**0.0379**	**0.9022**	**0.7633**	**0.0665**	**0.7948**	**0.9050**	**0.0369**	**0.8915**

The results of the top are shown in bold

Table 4 Extended research on the network structure [27]

Methods	ECSSD			MSRA-B			DUT-OMRON			HKU-IS		
	$F_\beta \uparrow$	MAE↓	S-measure↑	$F_\beta \uparrow$	MAE↓	S-measure↑	$F_\beta \uparrow$	MAE↓	S-measure↑	$F_\beta \uparrow$	MAE↓	S-measure↑
Baseline [27]	0.8431	0.1008	0.7790	0.8331	0.0863	0.7869	0.6490	0.1101	0.6513	0.8669	0.0615	0.8309
Baseline + hybrid upsampling	0.8717	0.0837	0.8275	0.8767	0.0605	0.8567	0.7257	0.0859	0.7317	0.8889	0.0466	0.8743
Baseline + hybrid upsampling + hybrid loss	**0.8847**	**0.0722**	**0.8508**	**0.8867**	**0.0549**	**0.8693**	**0.7319**	**0.0853**	**0.7413**	**0.8957**	**0.0422**	**0.8838**

The results of the top are shown in bold

Table 5 Extended research on the network structure [19]

Methods	ECSSD			MSRA-B			DUT-OMRON			HKU-IS		
	$F_\beta \uparrow$	MAE↓	S-measure↑	$F_\beta \uparrow$	MAE↓	S-measure↑	$F_\beta \uparrow$	MAE↓	S-measure↑	$F_\beta \uparrow$	MAE↓	S-measure↑
Baseline [19]	0.8949	0.0625	0.8748	0.8993	0.0478	0.8943	0.7418	0.0796	0.7704	0.8897	0.0480	0.8782
Baseline + hybrid upsampling	0.9059	0.0528	0.8963	0.9017	0.0440	0.9034	0.7559	0.0708	0.7934	0.9109	0.0330	0.9166
Baseline + hybrid upsampling + hybrid loss	**0.9065**	**0.0506**	**0.9002**	**0.9037**	**0.0412**	**0.9099**	**0.7622**	**0.0672**	**0.8065**	**0.9097**	**0.0322**	**0.9185**

The results of the top are shown in bold

these three methods, ours is closer to ground truth than the other two due to clearer boundaries.

4.5 Extended research on different network structure

Our hybrid upsampling and hybrid loss have been verified in the basic encoder-decoder network framework. In order to valuate their effect, different network frameworks are attempted including the network structure in [27] and NLDF [19]. Tables 4 and 5 show the effect of hybrid upsampling method and hybrid loss computing on these networks. They verify that hybrid upsampling and hybrid loss computing play an important role in improving the performance of baseline models. Baseline models combined with hybrid upsampling and hybrid loss achieve the best performance.

5 Conclusion

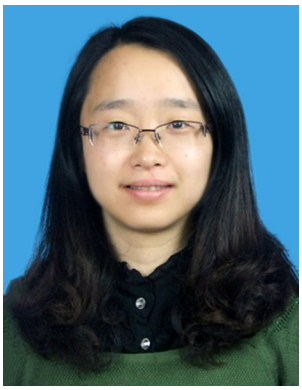
In this paper, we propose an end-to-end deep network for salient object detection based on hybrid upsampling method and hybrid loss computing. Hybrid upsampling method can not only combine original feature with upsampling feature, but also have a larger receptive field by using atrous convolution operation with different rate. Hybrid loss function composed of cross-entropy loss and area loss function can reduce the errors between the saliency map and the ground truth further. A fully connected CRF model can be optionally incorporated to further improve spatial coherence and contour localization. Experimental results demonstrate that our deep model can outperform the state-of-the-art method.

Acknowledgements We thank Prof. Ming-ming Cheng from Nankai University for providing the codes of all evaluation metrics and result saliency maps. We further thank all anonymous reviewers for their valuable comments. This research is supported by National Natural Science Foundation of China (61602004).

References

1. Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., et al.: Tensorflow: a system for large-scale machine learning. In: 12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16), pp. 265–283 (2016)
2. Achanta, R., Hemami, S., Estrada, F., Susstrunk, S.: Frequency-tuned salient region detection. In: 2009 IEEE Conference on Computer Vision and Pattern Recognition CVPR 2009, pp. 1597–1604 (2009)
3. Borji, A., Sihite, D.N., Itti, L.: Salient Object Detection: A Benchmark. Springer, Berlin (2012)
4. Chen, H., Li, Y.: Three-stream attention-aware network for RGB-D salient object detection. IEEE Trans. Image Process. (2019). https://doi.org/10.1109/TIP.2019.2891104
5. Chen, H., Li, Y., Su, D.: Multi-modal fusion network with multi-scale multi-path and cross-modal interactions for RGB-D salient object detection. Pattern Recognit. **86**, 376–385 (2019)
6. Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Deeplab: semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell. **40**(4), 834–848 (2018)
7. Cheng, M.M., Zhang, G.X., Mitra, N.J., Huang, X., Hu, S.M.: Global contrast based salient region detection. In: Computer Vision and Pattern Recognition, pp. 409–416 (2011)
8. Cong, R., Lei, J., Fu, H., Huang, Q., Cao, X., Hou, C.: Co-saliency detection for RBGD images based on multi-constraint feature matching and cross label propagation. IEEE Trans. Image Process. **PP**(99), 1–1 (2018)
9. Fan, D.P., Cheng, M.M., Liu, Y., Li, T., Borji, A.: Structure-measure: a new way to evaluate foreground maps. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 4548–4557 (2017)
10. Fan, D.P., Cheng, M.M., Liu, J.J., Gao, S.H., Hou, Q., Borji, A.: Salient objects in clutter: bringing salient object detection to the foreground. In: European Conference on Computer Vision, pp. 196–212. Springer (2018)
11. Ilg, E., Mayer, N., Saikia, T., Keuper, M., Dosovitskiy, A., Brox, T.: Flownet 2.0: evolution of optical flow estimation with deep networks. In: IEEE Conference on Computer Vision and Pattern Recognition (2017)
12. Jiang, P., Ling, H., Yu, J., Peng, J.: Salient region detection by UFO: uniqueness, focusness and objectness. In: IEEE International Conference on Computer Vision, pp. 1976–1983 (2013)
13. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected CRFs with Gaussian edge potentials. In: Advances in Neural Information Processing Systems, pp. 109–117 (2011)
14. Lee, G., Tai, Y.W., Kim, J.: Deep saliency with encoded low level distance map and high level features. In: Computer Vision and Pattern Recognition, pp. 660–668 (2016)
15. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: Computer Vision and Pattern Recognition, pp. 5455–5463 (2015)
16. Li, G., Yu, Y.: Deep contrast learning for salient object detection. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 478–487 (2016)
17. Li, X., Zhao, L., Wei, L., Yang, M.H., Wu, F., Zhuang, Y., Ling, H., Wang, J.: Deepsaliency: multi-task deep neural network model for salient object detection. IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc. **25**(8), 3919 (2016)
18. Liu, T., Sun, J., Zheng, N.N., Tang, X., Shum, H.Y.: Learning to detect a salient object. In: 2007 IEEE Conference on Computer Vision and Pattern Recognition CVPR '07, pp. 1–8 (2007)
19. Luo, Z., Mishra, A., Achkar, A., Eichel, J., Li, S., Jodoin, P.M.: Non-local deep features for salient object detection. In: Computer Vision and Pattern Recognition, pp. 6593–6601 (2017)
20. Mochizuki, I., Toyoura, M., Mao, X.: Visual attention prediction for images with leading line structure. Vis. Comput. **34**(6–8), 1031–1041 (2018)
21. Pan, J., Canton, C., Mcguinness, K., O'Connor, N.E., Torres, J., Sayrol, E., Giro-I-Nieto, X.: Salgan: visual saliency prediction with generative adversarial networks. (2017) arXiv preprint arXiv:1701.01081
22. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. Comput. Sci. (2014) arXiv preprint arXiv:1409.1556
23. e Souza, M.R., Pedrini, H.: Motion energy image for evaluation of video stabilization. Vis. Comput. pp. 1–13 (2017)
24. Wang, H., Dai, L., Cai, Y., Sun, X., Chen, L.: Salient object detection based on multi-scale contrast. Neural Netw **101**, 47–56 (2018a)
25. Wang, L., Lu, H., Ruan, X., Yang, M.H.: Deep networks for saliency detection via local estimation and global search. In: IEEE Confer-

ence on Computer Vision and Pattern Recognition, pp. 3183–3192 (2015)
26. Wang, T., Zhang, L., Lu, H., Sun, C., Qi, J.: Kernelized subspace ranking for saliency detection. In: European Conference on Computer Vision, pp. 450–466 (2016a)
27. Wang, W., Shen, J.: Deep visual attention prediction. IEEE Trans. Image Process. **PP**(99), 1–1 (2018)
28. Wang, W., Shen, J., Shao, L., Porikli, F.: Correspondence driven saliency transfer. IEEE Trans. Image Process. **25**(11), 5025–5034 (2016b)
29. Wang, W., Shen, J., Shao, L.: Video salient object detection via fully convolutional networks. IEEE Trans. Image Process. **27**(1), 38–49 (2018b)
30. Wang, W., Shen, J., Sun, H., Shao, L.: Video co-saliency guided co-segmentation. IEEE Trans. Circuits Syst. Video Technol. **28**(8), 1727–1736 (2018c)
31. Wang, W., Shen, J., Yang, R., Porikli, F.: Saliency-aware video object segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **40**(1), 20–33 (2018d)
32. Wei, Y., Wen, F., Zhu, W., Sun, J.: Geodesic saliency using background priors. In: European Conference on Computer Vision, pp. 29–42. Springer (2012)
33. Wenguan, W., Jianbing, S., Ling, S.: Consistent video saliency using local gradient flow optimization and global refinement. IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc. **24**(11), 4185 (2015)
34. Yan, B., Wang, H., Wang, X., Zhang, Y.: An accurate saliency prediction method based on generative adversarial networks. In: 2017 IEEE International Conference on Image Processing (ICIP), pp. 2339–2343. IEEE (2017)
35. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: Computer Vision and Pattern Recognition, pp. 1155–1162 (2013)
36. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3166–3173 (2013)
37. Zhang, D., Meng, D., Han, J.: Co-saliency detection via a self-paced multiple-instance learning framework. IEEE Trans. Pattern Anal. Mach. Intell. **39**(5), 865–878 (2016)
38. Zhang, P., Wang, D., Lu, H., Wang, H., Xiang, R.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: IEEE International Conference on Computer Vision, pp. 202–211 (2017a)
39. Zhang, P., Wang, D., Lu, H., Wang, H., Yin, B.: Learning uncertain convolutional features for accurate saliency detection. In: IEEE International Conference on Computer Vision, pp. 212–221 (2017b)
40. Zhu, W., Liang, S., Wei, Y., Sun, J.: Saliency optimization from robust background detection. In: Computer Vision and Pattern Recognition, pp. 2814–2821 (2014)



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

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

Peng Zhao is an Professor working with Key Laboratory of Intelligent Computing Signal Processing of Ministry of Education at Anhui University, China. Her research interests include Information Retrieval and Artificial Intelligence.