Edge-Guided Non-Local Fully Convolutional Network for Salient Object Detection

Zhengzheng Tu, Yan Ma, Chenglong Li, Jin Tang, and Bin Luo

***Abstract*— Fully Convolutional Neural Network (FCN) has been widely applied to salient object detection recently by virtue of high-level semantic feature extraction, but existing FCN-based methods still suffer from continuous striding and pooling operations leading to loss of spatial structure and blurred edges. To maintain the clear edge structure of salient objects, we propose a novel Edge-guided Non-local FCN (ENFNet) to perform edge-guided feature learning for accurate salient object detection. In a specific, we extract hierarchical global and local information in FCN to incorporate non-local features for effective feature representations. To preserve good boundaries of salient objects, we propose a guidance block to embed edge prior knowledge into hierarchical feature maps. The guidance block not only performs feature-wise manipulation but also spatial-wise transformation for effective edge embeddings. Our model is trained on the MSRA-B dataset and tested on five popular benchmark datasets. Comparing with the state-of-the-art methods, the proposed method performance well on five datasets.**

***Index Terms*— Salient object detection, edge guidance, non-local features, fully convolutional neural network.**

I. Introduction

SALIENCY detection is generally divided into two categories: salient object detection or eye fixation prediction, with the purpose of extracting the most predominant objects or predicting human eye attended locations corresponding to informative regions in an image. With the rapid development of deep learning, saliency detection has made great progress in recent years. As a preprocessing step, saliency detection is helpful for many applications in the computer vision tasks, such as image segmentation [1]–[3], scene classification [4], visual tracking [5]–[7], and person re-identification [8].

The saliency of each region is usually defined as the weight of the distance between the region and background regions of the the image in most traditional methods, and often based on hand-crafted features and some priors [9]–[11]. In recent years, with the success of Fully Convolutional Network (FCN) in the field of computer vision, deep learning methods has become a promising alternative to salient object detection. Existing methods based on the FCN [12], [12]–[19] have shown good performance, since the features obtained at deeper layers typically contain stronger semantic information and the global context-aware information, which are beneficial to locate salient regions even in complex scenes. FCN promotes saliency detection by applying multiple convolution layers and pooling layers to increase the receptive field and automatically extract high-level semantic information, which plays an import role in this field. These continuous convolution layers and pooling operations gain large receptive fields and high representation ability but reduce the size of feature maps and lose the spatial structure, and deteriorate the edges of the salient objects. This is useful for some high-level tasks like classification and recognition, but unfortunately, it reduces the accuracy of low-level tasks which usually require precise pixel activation, such as salient object boundaries. Some recent works notice this problem and propose some edge-aware saliency detection methods. For example, Zhang *et al*. [17] embed edge-aware feature maps achieved by shadow layers into the deep learning framework. Mukherjee *et al*. [20] construct a Bayesian probabilistic edge mapping which uses low-order edge features to assign a saliency value to the edge. However, boundary preserving still hasn't been solved well as lack of specific guidance for the boundary area. As shown in Fig. 1(Without), the predictions around object boundaries are inaccurate as the essential fine details are lost caused by repeated strides and pooling operations. In a word, existing FCN-based methods still have the drawback that they are difficult to maintain spatial structure like clear boundaries.

To handle this problem, we propose a novel method based on an edge-guided non-local FCN, named ENFNet, for salient object detection. We take the FCN model in [19] as the baseline of our ENFNet, and design an edge guidance block to perform feature-wise manipulation and spatial-wise trans-

Manuscript received August 28, 2019; revised December 27, 2019 and February 16, 2020; accepted March 8, 2020. Date of publication March 16, 2020; date of current version February 4, 2021. This work was supported in part by the National Natural Science Foundation of China under Grant 61602006, Grant 61702002, Grant 61976003, and Grant 61976002, in part by the NSFC Key Projects in International (Regional) Cooperation and Exchanges under Grant 61860206004, in part by the Natural Science Foundation of Anhui Province under Grant 1808085QF187 and Grant 1908085QF264, in part by the Natural Science Foundation of Anhui Higher Education Institution of China under Grant KJ2019A0026, and in part by the Open fund for Discipline Construction, Institute of Physical Science and Information Technology, Anhui University. This article was recommended by Associate Editor G. Zhao. *(Corresponding author: Chenglong Li.)*

Zhengzheng Tu, Yan Ma, Jin Tang, and Bin Luo are with the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei 230601, China (e-mail: zhengzhengahu@163.com; m17856174397@163.com; tangjin@ahu.edu.cn).

Chenglong Li is with the Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei 230601, China, and also with the Institute of Physical Science and Information Technology, Anhui University, Hefei 230601, China (e-mail: lcl1314@foxmail.com).

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

Digital Object Identifier 10.1109/TCSVT.2020.2980853

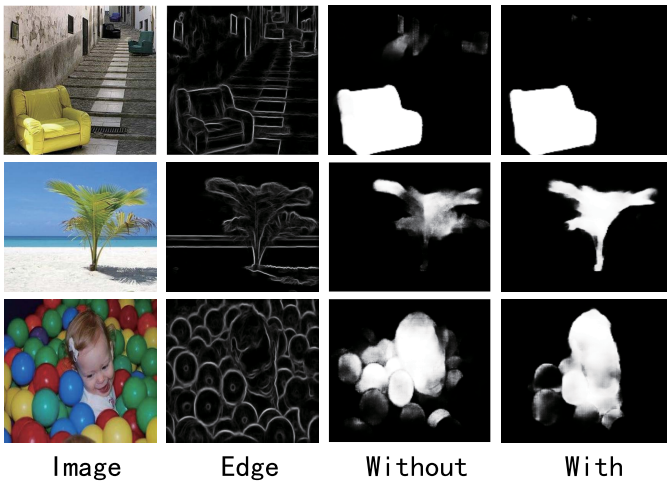


Fig. 1. Illustration of the importance of edge information to saliency detection. Image: input images. Edge: edge maps. Without: saliency results without edge embeddings from NLDF. With: saliency results with edge embeddings from ENFNet.

formation for effective edge embeddings. In a specific, given the input image, we first extract the edge maps using a EdgeNet [21], and then employ the baseline FCN to perform salient object detection. To embed the edge prior knowledge into the baseline FCN, we propose to use the spatially feature-wise affine transform algorithm [22] to guide feature learning. An example of edge-guided saliency map generation by our method is shown in Fig. 1.

In conventional models, the local contrast is widely used and plays an important role as salient objects should be different from background areas. While in deep models, the global context is very useful to model salient objects in the full image and complementary to the local contrast information. To fully utilize both local contrast and global context, we design a hierarchical edge-guided non-local structure in our network. In particular, features of each convolutional layer in FCN are first transformed into edge-guided features by edge prior knowledge, and then we employ the edge-guided features to generate local contrast features. The global context features are computed by FCN and combined with the edge-aware local contrast features to produce the final saliency map. We compare the proposed approach with many start-of-the-art saliency detection methods on the five benchmark datasets including HKU-IS [23], PASCAL-S [24], DUT-OMRON [10], ECSSD [11] and SOD [25], and our method outperforms most compared state-of-the-art methods under all evaluation metrics.

In short, the major contributions of this work are summarized as follow:

- We propose a novel edge-guided non-local fully convolutional network for salient object detection. The proposed network is able to embed the detailed edge information in a hierarchical manner and thus generate high-quality boundary-aware saliency maps.
- We design an edge guidance block to incorporate the edge prior knowledge in the non-local feature learning framework. The designed block is simple yet effective and thus yields a state-of-the-art performance with a little impact on execution time.
- Extensive experiments have proved that the proposed network is effective in producing good saliency maps with clear boundaries, and outperforms most compared state-of-the-art methods in all metrics.

II. Related Work

We will review two aspects of salient object detection approaches that are based on deep learning techniques and edge prior information.

A. Deep Learning Based Methods

Salient object detection can be regarded as a pixel-wise classification problem. Although traditional salient object detection methods have their superiority, including no need training and simplicity, their overall performances are not as good as most deep learning methods. Deep learning based methods have achieved great improvement in salient object detection as they combine local and deep features and can be trained end-to-end. In the state-of-the-art models of convolutional neural network (CNN), feature selection between salient and non-salient regions is automatically accomplished by gradient descent algorithm. These models consist of convolutional and pooling layers that are aggregated into the softmax layer, which estimates the probability that each pixel belongs to the object. Recently, these methods [12], [12]–[18], [23], [26], [27] have made a great process, which can be divided into two categories,that are based on the region of image and based on the fully convolutional network (FCN) respectively.

1) Region-Based Methods: The region-based deep learning methods use image patches as the basic processing units for saliency detection. For example, Li and Yu [23] adopt multi-scale features to obtain contextual information,then use a classifier to estimate the saliency value of each pixel in the image. Xu *et al*. [28] propose a deep network to fuse complementary information from multiple CNN side outputs, and the feature integration is obtained through continuous Conditional Random Fields(CRFs). And in [29], they utilize both Pyramid spatial pooling and multi-scale feature fusion to detect salient object regions with different sizes. Zhao *et al*. [30] propose a multi-context deep CNN framework benefiting from the global context of salient objects, as global context is conducive to modeling saliency of the image, while local context is conducive to estimating saliency of the region with rich features. Li and Yu [14] propose a deep contract network to combine a piece-wise stream and a pixel-level stream for saliency detection. And Wang *et al.* [13] propose to train two deep neural networks to integrate global search and local estimation for salient object detection. The work in [12] employs one branch to extract high-level features from VGG-net and another branch to obtain low-level features like color histogram. The fully-connected layers in CNNs is always to predicted the saliency score of every region [31]. These region-based approaches obviously improve the results of the methods using handcrafted features, however they lose some valuable spatial information. In addition, the region-based methods are time-consuming, as the network has to run many times.

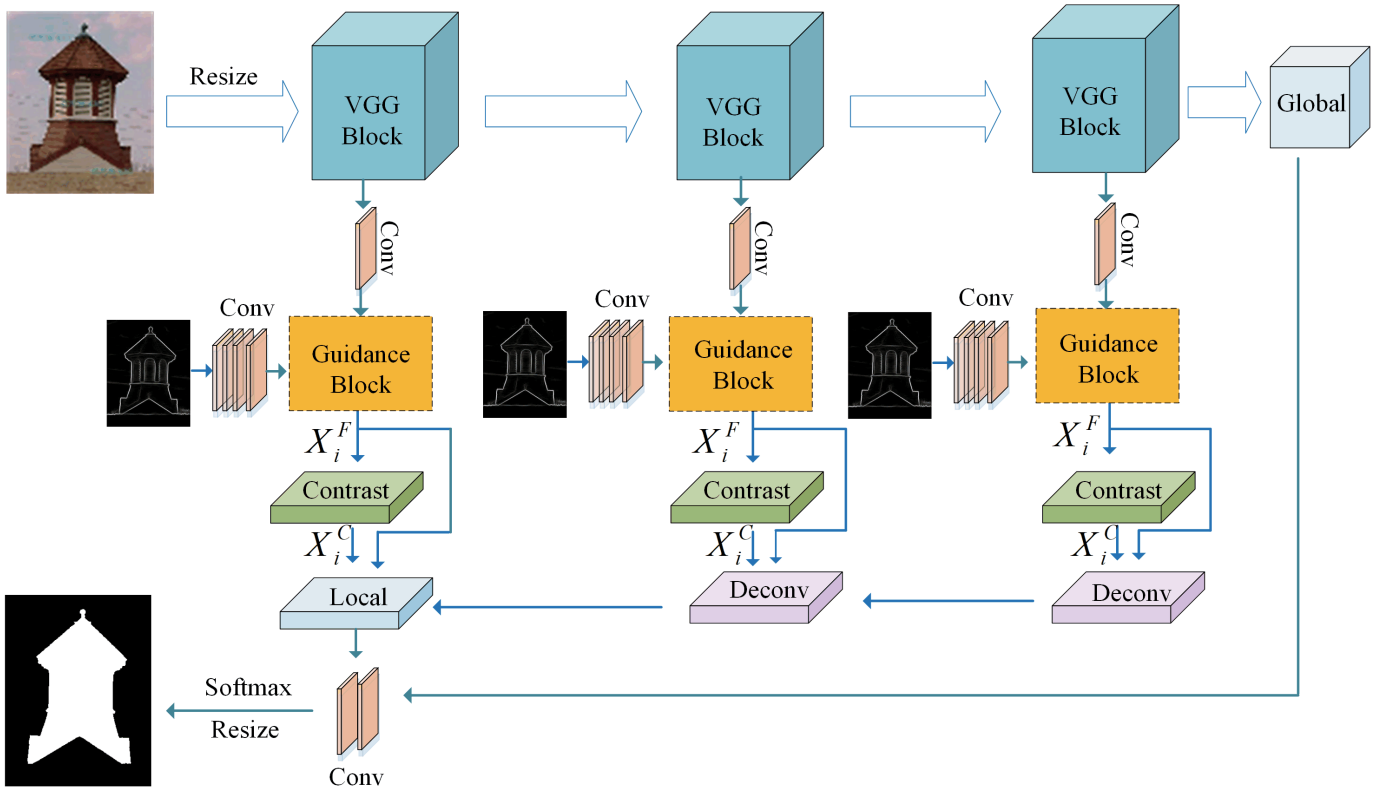


Fig. 2. Overall architecture of our proposed model. Our architecture is based on VGG-16 [32] for better comparison with previous CNN-based methods. Herein, only three convolution blocks are listed for clarity. The dotted-line part is an edge guidance block, where the edge features are regarded as the condition information on feature learning.

2) FCN-Based Methods: To improve the efficiency of CNN methods, people utilize FCN to generate a pixel-wise prediction. The FCN-based methods abandon the fully connected layer in CNN. They can increase the computational efficiency, but lost the spatial information. For example, Li and Yu [14] make use of segment-level spatial pooling stream and pixel-level fully convolutional stream to generate saliency estimation. The work in [26] designs a deep recurrent FCN to incorporate the coarse estimations as saliency priors and refines the generated saliency stage by stage. Xie and Tu [33] propose a new edge detection algorithm that can automatically learn rich hierarchical representations, and they address two critical issues including holistic image training and multi-scale feature learning. RAS [34] uses residual learning to learn residual features to refine saliency map, and proposes the reverse attention to guide these residual features. Inspired by HED [33], which builds skip-connections to employ multi-scale deep features for edge detection. Then, Hou *et al.* [16] propose dense short connections to skip-layers within the HED architecture to obtain rich multi-scale features for salient object detection, as the edge detection task is easier than saliency detection since it does not rely on much high-level semantic information. Another work by SRM [35] proposes a pyramid pooling model and a stage-wise refinement model to integrate both global and local context information. Zhang *et al.* [17] aggregate high level and low level features by concatenating feature maps directly, these multi-level features are beneficial to recover local details and locate salient objects. In addition, many methods attempt to find a way to fuse multi-scale features for better distinguishing salient and non-salient regions. However, without edge information, they can roughly detect the objects but cannot uniformly highlight the entire objects, and the estimated saliency maps suffer from the blur boundary.

B. Exploiting Edge Information

To solve above-mentioned blurred boundary problem, some methods attempt to take advantage of edge information for generating saliency maps with clear boundaries. Mukherjee *et al.* [20] construct a Bayesian probabilistic edge mapping and assign a salient value to the edge by using low-order edge features, then they learn a conditional random field to effectively combine these features with object/non-object labels for edge classification. The work in [36] and [37] use edge information for detecting salient regions. For example, in [36], they exploit a pre-trained edge detection model to detect the edge of objects. Then, based on the detected edge, they segment the image into regions, and generate saliency map of every region through a mask-based Fast R-CNN [38]. The model in [39] uses three different labels including salient objects, salient object boundaries and background. And in [40], they add an edge detection branch into the pooling network, through the joint training saliency model and edge detection, the details of salient objects are further sharpened. To enhance the accuracy of saliency boundary, the model [39] takes the extra hand-craft edge feature as a complementary to preserve

edge information effectively. Guan *et al*. [41] propose an edge detection stream to combine multiple side outputs together through concatenation and use a fusion layer that a 1×1 convolution to get the unified output. With the edge information added, these above models achieve good performance in preserving the boundary of salient object. However, these methods usually use edge losses to preserve the edges of salient objects or fuse shallow features and edge features for simple feature aggregation. To make better use of edge information, from a global and local perspective, we not only add edge features in shallow layers but also in deep layers. At the same time, our edge guidance block which includes a series of operations on edge features can achieve feature-wise affine transform, and retain spatial structure information more effectively. Through using parallel edge features to guide saliency detection, we could obtain more accurate saliency maps with clear object boundaries.

III. Edge-Guided Non-Local FCN

In this section, we will introduce the architecture structure of our proposed Edge-guided Non-local Fully convolutional Network (ENFNet). The overall structure of ENFNet is shown in Fig. 2.

A. Overview of Network Architecture

As discussed above, we combine global context information and local structure information with different resolution details, and integrate edge priors into different resolution features to better locate the boundary of the salient object. Existing methods always adopt the complicated backbone network [42] to obtain good features. To improve the efficiency, we adopt the VGG-16 network [32] as our backbone network like [19]. The image size of the input model is fixed to 352×352, and the output saliency map is 176×176. To obtain the saliency map with same size as the input image, we use the bilinear interpolation method to upsample feature maps. However, when the foreground and background have the same contrast, and the object has a complex background, the predicted salient results are often not so good, such as unclear boundaries. Therefore, to get good boundaries of salient objects, we design an edge guidance block to embed edge prior information into hierarchical feature maps for effective feature representations. The edge prior knowledge is beneficial to guide the results of saliency detection to possess good object boundaries as shown in Fig. 1.

As shown in Fig. 2, our ENFNet does not use the fully-connected layers because the task of saliency detection is focused on pixel-level prediction. We select the first five blocks of VGG-16 as side outputs, where each side output is followed by a convolution block, which is designed to obtain hierarchical multi-scale features $\{X_1, X_2, X_3, X_4, X_5\}$. Each convolution block contains two or three convolutional layers and a max pooling operation. The last block behind the backbone network contains three convolutional layers to computes global features X_G that is the global context of the image. The middle layers is the proposed edge guidance block, which is located on the right side of Fig. 2. Details of the edge guidance block are as shown in Table I. First, we extract edge maps using the existing method [21], and these edge maps X_i^E are used as input of the edge guidance block. Second, the edge features X_i^E and the hierarchical multi-scale features $X_i (i \in \{1, 2, 3, 4, 5\})$ are fused by the edge guidance block to generate edge-aware features X_i^F. Finally, the edge features X_i^F are through average pooling operation to obtain the contrast features X_i^C. Because the salient objects are related to global or local contrast, the contrast features capture the difference between each region and its neighbors. The last line of our model is a series of deconvolution blocks, and these deconvolution operation update the features maps from 11×11 to 176×176, for the convenience of concatenating feature maps (X_i^F, X_i^C) with different scales. The local blocks contain one convolution layer to compute the local feature X^L. At the last line, we employ two convolution layers to fuse the local and global features and then a softmax operation to obtain the saliency map.

TABLE I
Details of the Proposed Edge Guidance Block (EG Block)

Block	Layer	kernal	S	Pad	Output
CN	4 conv	3*3	1	Yes	176*176*128
EG Block-1	2 conv	3*3	1	Yes	176*176*128
EG Block-2	2 conv	3*3	2	Yes	88*88*128
EG Block-3	2 conv	3*3	2	Yes	44*44*128
EG Block-4	2 conv	3*3,5*5	2,4	Yes	22*22*128
EG Block-5	2 conv	5*5	4	Yes	11*11*128

B. Hierarchical Non-Local Structure

Some works show that lower layers in convolutional neural networks can capture rich spatial information, while higher layers encode object-level information but are invariant to the factors such as appearance and pose [43]. Generally speaking, the receptive field of lower layers is limited, it is unreasonable to require the network to perform dense prediction in the early stage. Performing feature extraction instead of feature classification at earlier stages, those extracted low-level features can provide spatial information for our final predictions at the bottom layer. Therefore, the multi-scale features are quite important to improve the performance of saliency detection. As shown in Fig. 2, these convolution blocks below in side outputs are processed by the first five convolutional blocks of VGG-16. The function of these convolution blocks is to obtain different scale features $\{X_1, X_2, X_3, X_4, X_5\}$. Each of convolution block uses the kernel size of 3×3 and outputs features with 128 channels.

1) Contrast Features: Saliency detection is related to the global or local contrast of foreground and background. In an image, salient objects are the foreground that highlights from the background. That is to say, salient features must be evenly distributed in the foreground, and the foreground and background regions are different. In order to capture such contrast information, we add contrast features related to each stage edge features X_i^F outputted by the edge guidance block. Each contrast feature is computed by subtracting X_i^F from

its local average, where the size of local neighbor region is 3×3. The average pooling operation that can reduce the errors caused by the variance of the estimated value due to the limited size of the neighborhood. Hence subtracting $AvgPool(X_i^F)$ can retain more foreground information, and salient objects are easier to detect:

$$X_i^c = X_i^F - AvgPool(X_i^F). \tag{1}$$

2) Deconvolution Features: After the last step, the sizes of the five contrast features is gradually decreasing. The first size of contrast feature X_1^c is 176×176, but the fifth size of contrast feature X_5^c is 11×11. To obtain the same size of final output 176×176, we need from back to front sequentially use five deconvolution (Deconv) layers to increase the sizes of precomputed feature maps X_i^F and X_i^C. At each Deconv blocks, we upsample the previous feature by the stride of 2. The result of the deconvolution feature map D_i is computed by combining the information of edge feature X_i^F and local contrast feature X_i^C. We upsample the feature D_i to obtain D_{i+1}. The Deconv operation is achieved by deconvolution layer with the kernel size of 5×5 and the stride of 2. The input of this Deconv layer is the concatenation of X_i^F, X_i^C and D_{i+1} and the channels number of D_{i+1} is sum of X_i^F and D_{i+1},

$$D_i = Deconv(X_i^F, X_i^C, D_{i+1}). \tag{2}$$

3) Local Features: The local block uses a convolution layer with a stride of 1 and a kernel size of 1×1 to obtain the final local features maps X^L:

$$X^L = Conv(X_1^F, X_1^C, D_2), \tag{3}$$

where the channel number of X_L is the sum of sizes of X_1^F and D_2, and the size of the local feature X_L is 176×176. For convenience to fusion with the global feature, we need again use the Deconv operation to increase the size of the local feature X^L from 176×176 to 352×352.

4) Global Features: A good salient object detection model can not only capture local features but also capture global features. Before assigning saliency to a single small region, the saliency model needs to acquire the global context information of the image. To achieve this purpose, we use three convolution layers after the last VGG block to computer global feature X_G. The first two convolution layers use the kernel size of 5×5, the last convolutional layer uses the kernel size of 3×3 and all convolutional layers have the same channel dimensions.

C. Edge Guidance Block

Five side outputs will generate five features with different scale, which have different channel number. To effectively embed the edge feature X_i^E into the feature X_i, we design a condition network to generate shared intermediate conditions in all layers for efficient computation, as shown in Fig. 3. These generated conditions are regarded as one of inputs of the edge guidance block. The inspiration of the edge guidance block comes from [22], which employs deep spatial feature transform to recover realistic texture in image super-resolution. And they adopt the possibility of semantic segmentation maps as the categorical prior, and the spatial feature transform layer is conditioned on semantic segmentation probability maps. Different from [22], as shown in Fig. 3, our edge guidance block is composed of two stages. In the first stage, our condition network uses four convolutional layers with the edge map as input to generate conditional features X_i^E, where each convolutional layer is with a kernel size of 3×3 and a stride of 1. The size of X_i^E is 176×176. Note that the condition network only produces coarse edge feature X_i^E. To better transfer edge features to next layer of the network and play a better guidance role, we propose to aggregate the edge guidance block into hierarchical feature maps. The edge guidance block uses two separate branches to output two features (γ, β) based on the conditional features. Then, we use (γ, β) to transform X_i into an edge-aware feature X_i^F as follows:

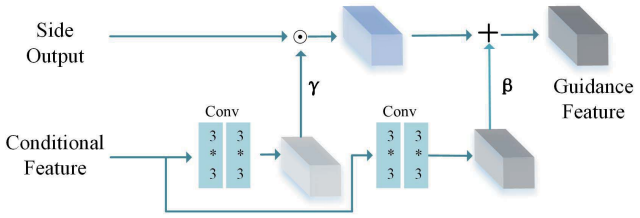


Fig. 3. Details and calculation flow of edge guidance block.

$$EGB(X_i^F|\gamma, \beta) = X_i \odot \gamma + \beta, \tag{4}$$

where $\odot$ represents the element-wise product operation and $+$ represents the element-wise addition operation. Since the spatial dimensions are preserved, the edge guidance block not only performs feature-wise manipulation but also spatial-wise transformation. As same as NLDF [19], we take five edge guidance blocks to obtain local hierarchical features, and utilize three convolution layers after the last VGG block to acquire the global context information of the image. Finally, we combine the local edge-related multi-scale features and global context information to realize our non-local fully convolutional network for salient object detection.

D. Loss Function

Saliency detection and image segmentation usually evolve into the optimization problem of non-convex energy functions.The energy function consists of data items and regularization items. A mathematical global model is the Mumford-Shah (MS) model [44], which transforms the image segmentation problem into solving the minimum value of energy function. By constructing the energy function, the curve is evolved under the driving of the minimum value of energy function and the contour curve gradually approaches the boundary of the object, and the object is finally segmented. In [19], they propose a supervised deep learning method, and the purpose of MS functional is to minimize IOU loss by maximizing the coincidence rate between predicted boundaries

and real boundaries:

$$F_{MS} \approx \underbrace{\sum_j \lambda_j \int_{v \epsilon \Omega_j} S_j(y(v), \hat{y}(v)) dv}_{cross\ entropy\ loss} + \underbrace{\sum_j \gamma_j (1 - IoU(C_j, \widehat{C}_j))}_{boundary\ loss} \quad (5)$$

where v is the pixel location, S_j is the total loss between ground truth (y) and predicted $(\hat{y})$ saliency map, and $\mathrm{IoU}(C_j, \widehat{C}_j)$ is the intersection over union between the true boundary and estimated boundary. The constant λ_j and γ_j are used to adjust the energy function according to data item and total boundary length.

1) Cross-Entropy Loss: We use two linear operators (W_G, b_G) and (W_L, b_L) to combine the local and global features. And the softmax function is used to compute the probability for each pixel of being salient or not:

$$\hat{y} = p(y(v) = s) = \frac{e^{W_L^s X_{L(v)} + b_L^s + W_G^s X_G + b_G^s}}{\sum_{s' \in \{0,1\}} e^{W_L^{s'} X_{L(v)} + b_L^{s'} + W_G^{s'} X_G + b_G^{s'}}} \quad (6)$$

where p is the probability for each pixel of being salient, v is the location of a pixel and s indicates that the pixel belongs to the foreground. The cross-entropy loss function is:

$$G_j(y(v), \hat{y}(v)) = -\frac{1}{N} \sum_{i=1}^{N} \sum_{s \in \{0,1\}} (y(v_i) = s)(\log(\hat{y}(v_i = s)) \quad (7)$$

where $y(v)$ is the ground truth and $\hat{y}(v)$ is the predicted saliency map.

2) IoU Boundary Loss: This inspiration comes from the applications of the IOU boundary loss in image segmentation [45], [46]. IOU boundary loss computes the error between the true boundary C_j and predicted boundary $\hat{C}_j$. The boundary pixels employ a Sobel operator and use a tanh activation function before, which estimates the gradient magnitude of saliency maps with a probability scope of [0,1]. The IOU boundary loss can be represented as:

$$IoU\,Loss = 1 - \frac{2 \mid C_j \cap \hat{C}_j \mid}{\mid C_j \mid + \mid \hat{C}_j \mid} \quad (8)$$

IV. Experiments

This section will present the details of our evaluation settings, comparison with the state-of-the-art methods and impact of the proposed edge guidance block on the detection performance.

A. Experimental Setup

1) Evaluation Datasets: The proposed model is evaluated on five public benchmark datasets including HKU-IS [23], PASCAL-S [24], DUT-OMRON [10], ECSSD [11] and SOD [25]. Their details are as follows.

- *HKU-IS:* This dataset contains 4447 high quality pixel labeled images,most of which have low contrast and many salient objects. It also contains many independent salient objects or objects touching the image boundary.
- *PASCAL-S:* This dataset has 850 natural images generated from the PASCAL VOC [55] segmentation challenge. The ground truths labeled by 12 experts includes both pixel-wise labeled saliency and eye fixation.
- *DUT-OMRON:* This is a large dataset with 5168 high quality images. Each image in this dataset has one or more salient objects and a cluttered background. Therefore, this dataset is more challenging for saliency detection.
- *ECSSD:* This dataset contains 1000 natural and complex images picked from the internet,all images have been labeled in pixel-level as ground truth.
- *SOD:* This dataset has 300 images, and it was originally designed for image segmentation. This dataset is challenging because many images have low contrast, contain multiple salient objects, or the object touches boundaries of image.

2) Implementation: We implement our network on one Nvidia GTX 1070Ti GPU and in Tensorflow [56]. We choose NLDF [19] as our baseline, the pre-trained VGG-16 model [32] is used to initialize the weights in first five VGG Blocks. We initialize randomly all weights in convolution and deconvolution layers with a non-zero constant (0.01), and initialize the biases to 0. The learning rate is $10e{-}5$ and λ_j, γ_j in (5) were set to 1. We use the Adam optimizer [57] to train our model. For a fair comparison, our network and the baseline NLDF [19] adopt the same training dataset. MSRA-B [58] dataset contains 5000 images and most of the images have one salient object. As same as the NLDF [19], we take horizontal flipping on MSRA-B [58] for data augmentation, and obtain an augmented dataset with twice size of the original MSRA-B. We train our network on this augmented dataset.

All training images are resized to 352×352. We adopt the cross entropy loss and the IOU boundary loss to optimize our network. The entire training process requires 10 epochs with a single image batch size, and needs approximately ten hours to complete all training. After training, our method maintains a fast runtime of 0.08 second per image.

3) Evaluation Criteria: The performances of our network are evaluated by three different evaluation metrics, including F-measure score, mean absolute error(MAE) [59] and Precision-recall(PR) curves. With continuous intensity values normalized to the range of 0 to 255, we compute the corresponding binary map, then compute the precision /recall pairs of all binary maps of the dataset. The PR curve illustrates the average precision and recall rate of saliency maps at different thresholds, and the F-measure is a harmonic mean of precision and recall rate. F-measure is represented as:

$$F_\beta = \frac{(1 + \beta^2) \cdot Precision \cdot Recall}{\beta^2 \cdot Precision \cdot Recall} \quad (9)$$

where $\beta^2 = 0.3$ which emphasizes precision over recall as same as [60]. The mean absolute error (MAE) is to measure the average difference between estimated saliency map S and

TABLE II

MAXIMUM F-MEASURE (LARGER IS BETTER) AND MAE (SMALLER IS BETTER) OF DIFFERENT SALIENCY DETECTION METHODS ON FIVE RELEASED SALIENCY DETECTION DATASETS, THE THREE COLORS OF RED, GREEN AND BLUE RESPECTIVELY REPRESENT THE FIRST THREE METHODS WITH BETTER PERFORMANCE

Methods	HKU-IS		DUT-OMRON		PASCAL-S		ECSSD		SOD	
	max F_β	MAE	max F_β	MAE	max F_β	MAE	max F_β	MAE	max F_β	MAE
ENFNet	0.915	0.040	0.779	0.065	0.843	0.095	0.921	0.051	0.831	0.134
JointCRF [47]	0.920	0.039	0.801	0.057	0.867	0.085	0.927	0.049	0.811	0.137
RAS [34]	0.913	0.046	0.786	0.062	0.836	0.106	0.921	0.056	0.822	0.143
NLDF [19]	0.904	0.047	0.754	0.084	0.833	0.099	0.905	0.061	0.814	0.140
Amulet [17]	0.899	0.050	0.743	0.098	0.839	0.099	0.915	0.059	0.778	0.156
DHS [48]	0.892	0.052	-	-	0.832	0.095	0.907	0.059	0.812	0.135
DSS [16]	0.900	0.050	0.760	0.074	0.832	0.104	0.908	0.062	0.807	0.145
DS[49]	0.864	0.078	0.745	0.120	0.766	0.176	0.882	0.122	0.771	0.199
WSS [50]	0.858	0.079	0.689	0.110	0.782	0.141	0.856	0.103	0.759	0.181
UCF [51]	0.888	0.061	0.730	0.120	0.825	0.115	0.903	0.069	0.796	0.159
LEGS [13]	0.770	0.118	0.669	0.133	0.756	0.157	0.827	0.118	0.707	0.215
DCL [14]	0.892	0.054	0.733	0.084	0.815	0.113	0.887	0.072	0.795	0.142
MDF [23]	0.861	0.076	0.694	0.092	0.764	0.145	0.832	0.105	0.745	0.192
MCDL [30]	0.808	0.092	0.701	0.089	0.745	0.146	0.837	0.101	0.704	0.194
wCtr [52]	0.735	0.138	0.588	0.171	0.664	0.199	0.726	0.165	0.639	0.231
DRFI [53]	0.777	0.144	0.664	0.150	0.696	0.210	0.782	0.170	0.672	0.242
MR [10]	0.709	0.174	0.610	0.187	0.653	0.232	0.742	0.186	0.629	0.274

ground truth G:

$$MAE = \frac{1}{W \times H} \sum_{x=1}^{W} \sum_{y=1}^{H} | S(x, y) - G(x, y) | \qquad (10)$$

where W and H is the width and height of a given image.

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

To verify the effectiveness of our experiments, we compare our method with other state-of-the-art ones including three conventional methods (MR [10], DRFI [53], wCtr [52]) and thirteen deep learning based methods (MCDL [30], MDF [23], DCL [14], DSS [16], LEGS [13], UCF [51], WSS [50], Amulet [17], NLDF [19], RAS [34], DS [49], DHS [48], JointCRF [47]. It is worth mentioning that Amulet [17] also utilizes edge information. DHS [48] uses DUT-OMRON dataset to train the model, therefore we test DHS [48] on other four datasets. Note that since training set in the HKU-IS dataset is used for training MDF [23], we only calculate the evaluation metrics on the test set in HKU-IS. Because MDF [23] provides 200 saliency maps on SOD dataset, we also uses these saliency maps for evaluating other methods on SOD. For fair comparison, we use the implementations with suggested parameter settings and the saliency maps provided by authors.

1) Quantitative Evaluation: We compare the proposed method with others in terms of F-measure scores, MAE scores, and PR-curves, and the quantitative comparison results on five public benchmarks verify the effectiveness of our method, as shown in Fig. 4 and Table II. We choose NLDF [19] as our baseline which does not contain the edge guidance block, and our ENFNet includes five edge guidance blocks. Our ENFNet outperforms the baseline method NLDF on five public benchmarks. It is easy to see that the F-measure value of our method increases 2.5% and 1.0% over the method NLDF on the two most challenging datasets DUT-OMRON and PASCAL-S respectively. MAE value of our method decreases 1.9% and 0.4% over the baseline method NLDF on DUT-OMRON and PASCAL-S respectively. Therefore, the experimental results show that our edge guidance is effective. The comparison results also show that our approach is clearly superior to these state-of-the-art methods except for latest JointCRF [47] on HKU-IS, PASCAL-S, DUT-OMRON and ECSSD in terms of F-measures and MAE scores, as JointCRF [47] proposes a cascade CRFs architecture with CNN to jointly refine deep features and predicted maps at each scale, which is more robust for complex backgrounds. So JointCRF [47] has an obvious effect on DUT-OMRON as shown in Fig. 4, since DUT-OMRON has more images with cluttered background. However, our method achieves much better results on SOD dataset than JointCRF [47], as SOD has lots of images with low contrast or multiple salient targets, our ENFNet detects the salient objects with accurate boundaries. RAS [34] uses residual learning to learn residual features to refine saliency map, and further employs the reverse attention to guide these residual features. With the help of reverse attention, RAS [34] can discover the missing parts in salient objects, especically for the objects with cluttered backgrounds. Therefore, RAS [34] performs well on DUT-OMRON dataset, each image in that has more than one salient objects and cluttered background. In DHS [48], they makes a coarse global prediction by automatically learning various global saliency cues, then adopt a hierarchical recurrent convolutional neural network to integrate local context information. Compared with DHS [48], our ENFNet achieves better performance on all datasets. Compared with the top-level method DCL [14] with CRF-based post-processing to refine the resolution, our method still achieves the better performance.

2) PR Curves: In addition to the above results, we also give the PR curves on five datasets. As shown in Fig. 4, it can be seen that the PR curve (red) obtained by our method are prominent compared with other 16 methods. When the recall score is close to 1, our accuracy score is much higher than other most methods. This result shows that the false-positive

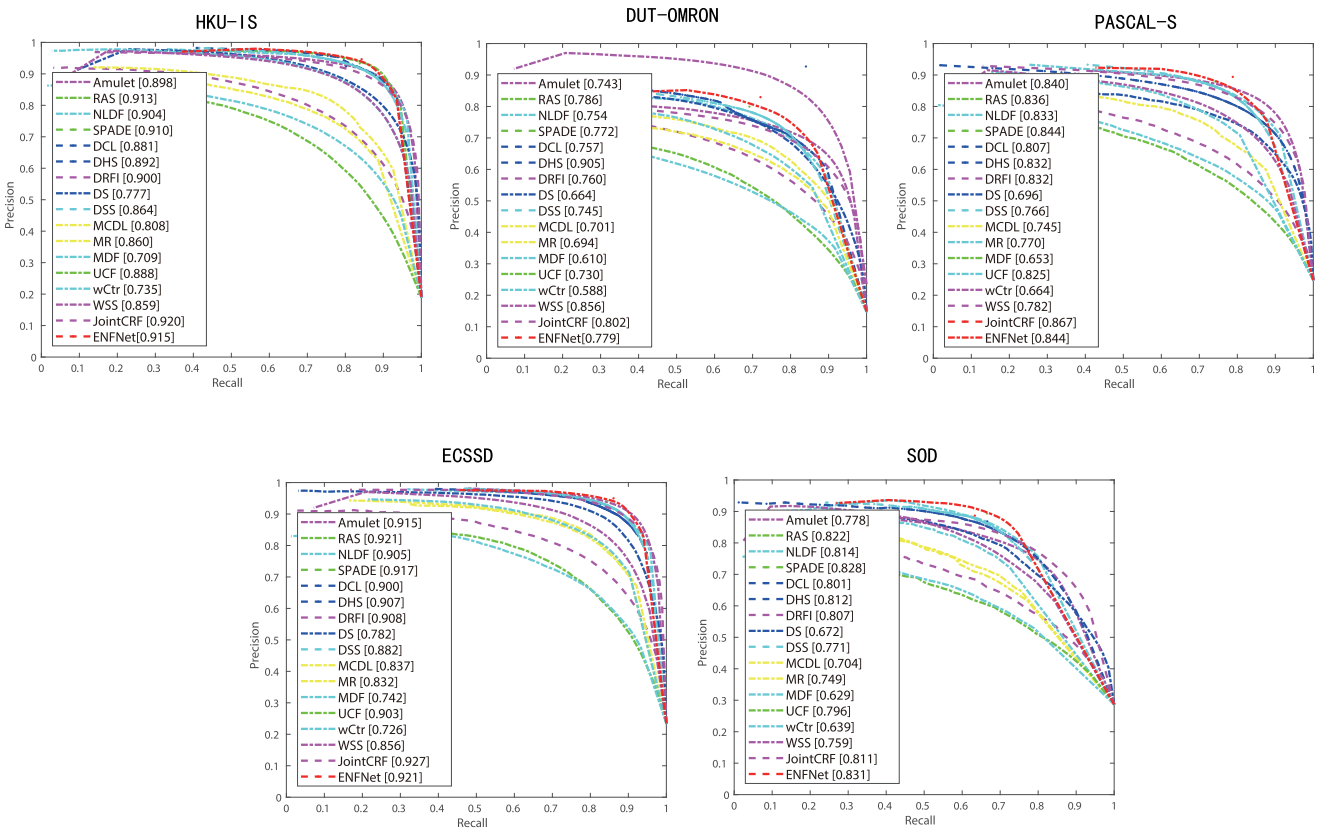


Fig. 4. Precision-recall curves of our model compared to MR [10], DRFI [53], wCtr [52], MCDL [30], MDF [23],DCL [14],DS [49], DSS [16], UCF [51], WSS [50], DHS [48], Amulet [17], ENFNet-SPADE [54], RAS [34], NLDF [19], JointCRF [47].

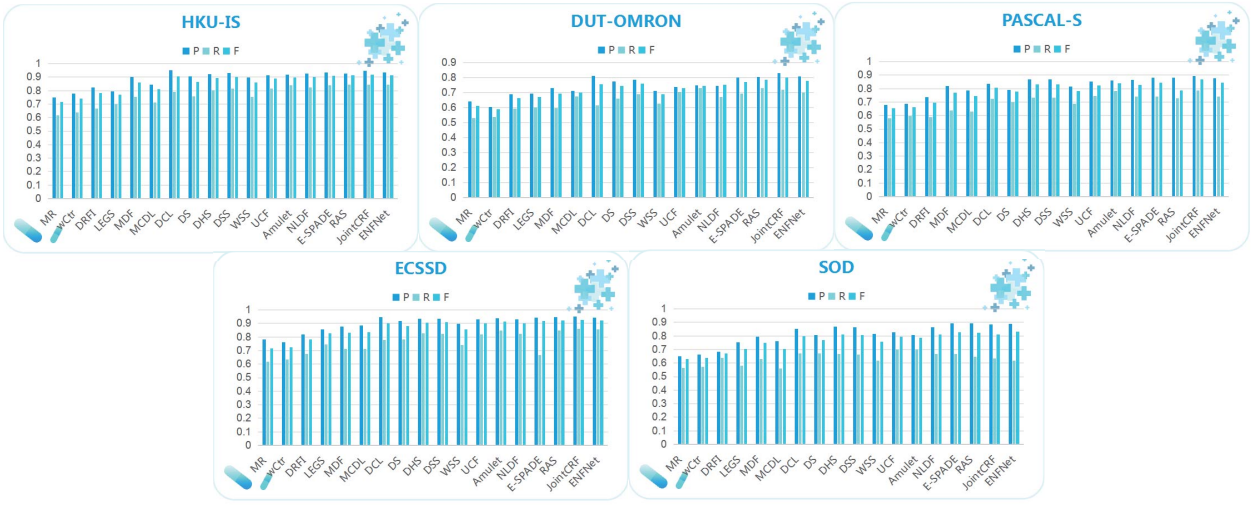


Fig. 5. For a better visual effect, we show the Precision, Recall and F-measure value in the form of a histogram compared to MR [10], DRFI [53], wCtr [52], MCDL [30], MDF [23], DCL [14], DS [49], DSS [16], LEGS [13], UCF [51], WSS [50], Amulet [17], DHS [48], ENFNet-SPADE [54], NLDF [19] and RAS [34],JointCRF [47].

rate of our saliency maps are lower than other methods. To express the superiority of our method more intuitively, we report the values of Precision, Recall and F-measure in the form of the histograms on five datasets and with 16 methods, as shown in Fig. 5.

3) Visual Comparisons: Fig. 6 provides a visual comparison of our approach with other methods. It can be seen our method generates more accurate saliency maps with clear boundaries. We select most representative images from each dataset to display the results. These images usually contain multiple salient objects and have complex background or unclear edges. From Fig. 6 we can see that our method obtains the best results which are much closer to the ground truth in various challenging scenarios. To be specific, with the help of edge guidance, the proposed method not only highlights the salient object regions clearly, but also generates the saliency maps with clear boundaries and consistent saliency values.

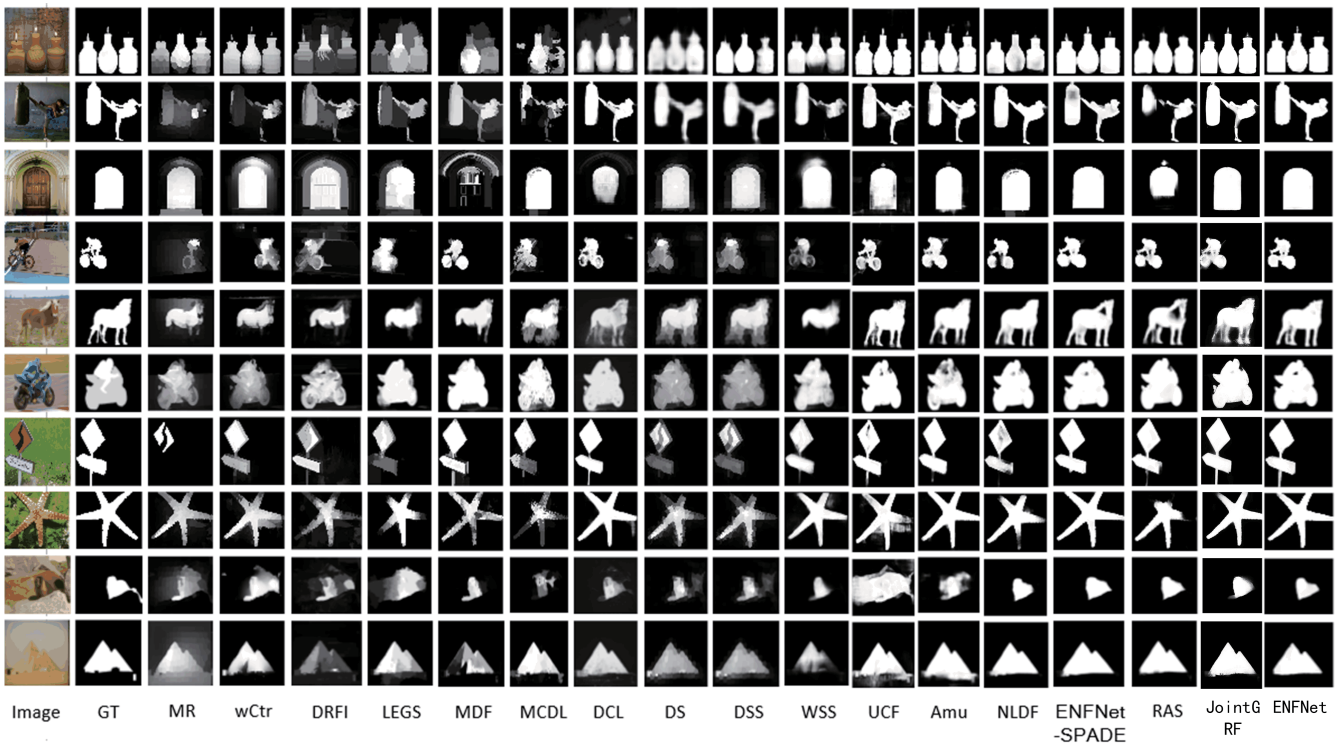


Fig. 6. Saliency maps produced by the MR [10], DRFI [53], wCtr [52], MCDL [30], MDF [23],DCL [14],DS [49], DSS [16], LEGS [13], UCF [51], WSS [50], Amulet [17], ENFNet-SPADE [54], RAS [34], NLDF [19] and JointCRF [47].

In addition, we provide some successful examples and failed examples of ENFNet compared with two top models as shown in Fig. 8. JointCRF [47] is the latest in these methods for comparison, its baseline is the backbone CNN based on the enhanced HED [33] structure, by adding message passing between deep features and predictions, and joint training CRF through back propagation to improve the overall performance of the model. As shown in the first line of Fig. 8, compared with RAS [34] and JointCRF [47], our method performs not well, as the first image has a cluttered background. In the second line of Fig. 8, for the second image with low contrast between background and object, similar to the fifth line, compared with JointCRF [47]and RAS [34], our method performs better. Contrast information is very important for judging salient objects, with the help of contrast module, our ENFNet can capture more contrast information which makes object standing out from its surrounding, as shown in the fourth and fifth line of Fig. 8. Edge information can guide the detection of salient objects, with the help of edge guidance, our ENFNet can detect the salient objects with the clear boundaries, as shown in the second line and fifth line of Fig. 8.

In short, compared with state-of-the-art methods, our ENFNet have these strengths. (1) Our edge guidance block performs not only feature-wise manipulation but also spatial-wise transformation, can refine boundaries of salient objects, therefore, our ENFNet can locate the salient object more accurately and further address the problem of blurry boundaries. (2) With the help of contrast module, our network can capture more contrast information which makes object standing out from its surrounding. The limitation of the proposed model is that when the information around the salient object is rich, not only the edge information but also some background information would be extracted. It would inevitably bring some noises to affect the overall performance of the model. Because our edge information is extracted by the existing edge detection model, our possible improvement in the future is to develop a robust edge detection model and integrate it into our framework to perform the end-to-end learning.

TABLE III

EFFECTS OF DIFFERENT NUMBER OF THE EDGE GUIDANCE BLOCKS (EGB) ON THE PERFORMANCE

Datasets	Five EGB		Three EGB		Zero EGB	
	max F_β	MAE	max F_β	MAE	max F_β	MAE
HKU-IS	**0.915**	**0.040**	0.906	0.046	0.904	0.047
DUT-OMRON	**0.779**	**0.065**	0.772	0.077	0.754	0.084
PASCAL-S	**0.843**	**0.095**	0.839	0.098	0.833	0.099
ECSSD	**0.921**	**0.051**	0.913	0.057	0.905	0.061
SOD	**0.831**	**0.134**	0.828	0.136	0.814	0.140

C. Impact of Edge Guidance Block

As shown in Fig. 6, visualization results of NLDF [19] without edge guidance block are not as good as our ENFNet. Because our method takes advantage of edge information, we also compare it with an edge-related method Amulet [17] in TABLE III and Fig. 4, in which our method also achieves better performance. To further demonstrate the effect of edge

TABLE IV
EFFECTIVENESS OF THE MAJOR COMPONENTS IN OUR METHOD, THE BOLD BLACK FONT REPRESENTS THE METHOD WITH BEST PERFORMANCE

Methods	HKU-IS		DUT-OMRON		PASCAL-S		ECSSD		SOD	
	max F_β	MAE	max F_β	MAE	max F_β	MAE	max F_β	MAE	max F_β	MAE
ENFNet	**0.915**	**0.040**	**0.779**	**0.065**	0.843	0.095	**0.921**	**0.051**	**0.831**	0.134
ENFNet-	0.907	0.044	0.765	0.072	0.838	0.097	0.916	0.055	0.823	0.138
ENFNet-SPADE [54]	0.910	0.043	0.772	0.071	**0.844**	**0.094**	0.917	0.053	0.828	**0.133**
NLDF [19]	0.904	0.047	0.754	0.084	0.833	0.099	0.905	0.061	0.814	0.140

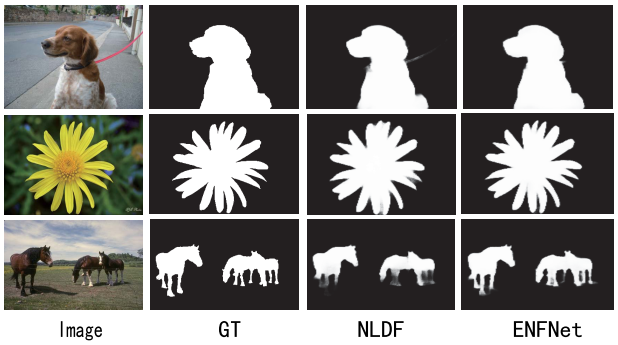


Fig. 7. Three examples showing effectiveness of edge information. Image: Input images. GT: Ground truths. NLDF: Saliency results of NLDF [19] without edge guidance. ENFNet: Saliency results of our ENFNet with edge guidance.

information,several visualization comparison with our baseline NLDF [19] are shown in Fig. 7. We can see that our approach can accurately localize the boundaries of most salient objects and produce more precise saliency maps, with the help of the edge guidance module. In addition, we also conduct some experiments to show the impact of the number of edge guidance blocks on saliency detection as shown in TABLE III. One experiment adopts three edge guidance blocks in first three layers, and another one removes all edge guidance blocks(that is our baseline NLDF [19]). From the results we can see that our method ENFNet(with five edge guidance blocks) is best, demonstrating the effectiveness of our design on the network.

D. Ablation Study

To verify the effectiveness of the proposed edge guidance block, we have done the ablation study on HKU-IS, DUT-OMRON, PASCAL-S, ECSSD and SOD datasets as shown in Table IV. First, compared with baseline NLDF [19] which has no edge guidance blocks, the inclusion of edge feature in our ENFNet leads to an increase in max F_β and decrease in MAE. Second, since SPADE [54] uses spatially-adaptive normalization similar to our guidance block, we replace the guidance block in ENFNet with spatially-adaptive normalization in SPADE, and list the comparison result named ENFNet-SPADE in Table IV, and the PR Curves and visual comparisons on five datasets are as shown in Fig. 4, Fig. 6 and Fig. 5. We can see that with the help of spatially-adaptive normalization, the experimental results of ENFNet-SPADE are better than the baseline method NLDF [19] on five datasets, and F-measure value of ENFNet-SPADE increases 1.8% and 1.1% over NLDF on two most challenging datasets DUT-OMRON and PASCAL-S respectively.

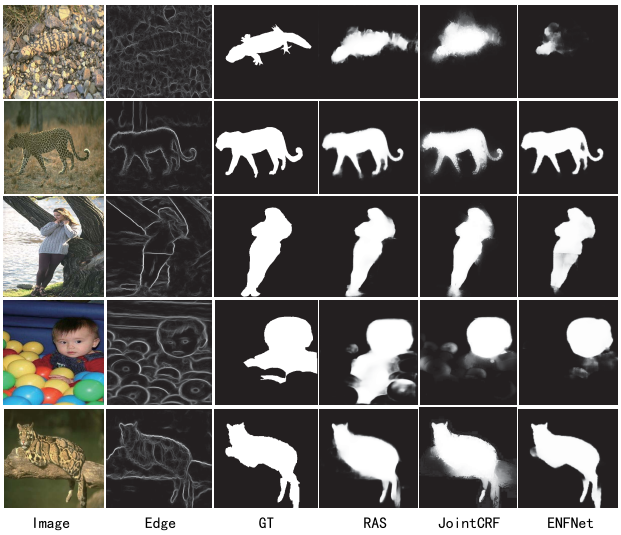


Fig. 8. Some successful examples and failed examples of ENFNet compared with RAS [34], JointCRF [47].

MAE value of ENFNet-SPADE decreases 1.3% and 0.5% over NLDF on DUT-OMRON and PASCAL-S respectively. But ENFNet-SPADE performs worse than our ENFNet on four datasets except for PASCAL-S dataset, which is generated from the PASCAL VOC segmentation challenge. On PASCAL-S datasets our ENFNet and ENFNet-SPADE obtain similar results. Therefore, the above ablation study certifies our edge guidance is effective. Finally, in order to show the effectiveness of the proposed non-local structure, we conduct a comparison experiment that we remove the part capturing global context information in ENFNet(named ENFNet-), listed in Table IV. The experimental results show that the non-local structure(in ENFNet) performs better for salient object detection than local structure(ENFNet-). Furthermore, compared with NLDF [19]that has the global structure and no edge guide blocks, the method ENFNet-, which has edge guide blocks and no global structure, also outperforms NLDF [19].

V. CONCLUSION

In this paper, we have proposed a novel Edge-guided Non-local FCN for salient object detection from the local and global perspective. We demonstrate that performing edge-guided feature learning with edge prior knowledge is beneficial to generating high-quality saliency results. Through the edge guidance block, the edge features are embedded into feature learning of our network, our method can preserve more accurate edge structure information. Experimental results demonstrate that

our proposed method consistently improves the performance on all five benchmarks and outperforms 16 state-of-the-art methods with different evaluation metrics. In future work, we will study other prior knowledge or information like semantic priors [61] or thermal infrared data [62] to help improving the performance of salient object detection, and also extend our framework to salient object detection in videos.

REFERENCES

[1] M. Donoser, M. Urschler, M. Hirzer, and H. Bischof, "Saliency driven total variation segmentation," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 817–824.

[2] C. Li, L. Lin, W. Zuo, W. Wang, and J. Tang, "An approach to streaming video segmentation with sub-optimal low-rank decomposition," *IEEE Trans. Image Process.*, vol. 25, no. 5, pp. 1947–1960, May 2016.

[3] C. Li, X. Wang, L. Zhang, J. Tang, H. Wu, and L. Lin, "Weighted low-rank decomposition for robust grayscale-thermal foreground detection," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 27, no. 4, pp. 725–738, Apr. 2017.

[4] Z. Ren, S. Gao, L.-T. Chia, and I. W.-H. Tsang, "Region-based saliency detection and its application in object recognition," *IEEE Trans. Circuits Syst. for Video Technol.*, vol. 24, no. 5, pp. 769–779, May 2014.

[5] A. Borji, S. Frintrop, D. N. Sihite, and L. Itti, "Adaptive object tracking by learning background context," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, Jun. 2012, pp. 23–30.

[6] C. Li, H. Cheng, S. Hu, X. Liu, J. Tang, and L. Lin, "Learning collaborative sparse representation for grayscale-thermal tracking," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5743–5756, Dec. 2016.

[7] C. Li, L. Lin, W. Zuo, J. Tang, and M.-H. Yang, "Visual tracking via dynamic graph learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 11, pp. 2770–2782, Nov. 2019.

[8] R. Zhao, W. Ouyang, and X. Wang, "Unsupervised salience learning for person re-identification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2013, pp. 3586–3593.

[9] F. Perazzi, P. Krahenbuhl, Y. Pritch, and A. Hornung, "Saliency filters: Contrast based filtering for salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 733–740.

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

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

[12] G. Lee, Y.-W. Tai, and J. Kim, "Deep saliency with encoded low level distance map and high level features," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 660–668.

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

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

[15] X. Shen and Y. Wu, "A unified approach to salient object detection via low rank matrix recovery," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2012, pp. 853–860.

[16] Q. Hou, M. Cheng, X. Hu, A. Borji, Z. Tu, and P. H. S. Torr, "Deeply supervised salient object detection with short connections," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3203–3212.

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

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

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

[20] P. Mukherjee, B. Lall, and S. Tandon, "Salprop: Salient object proposals via aggregated edge cues," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2017, pp. 2423–2429.

[21] C. L. Zitnick and P. Dollár, "Edge boxes: Locating object proposals from edges," in *Proc. IEEE Eur. Conf. Comput. Vis.*, 2014, pp. 391–405.

[22] X. Wang, K. Yu, C. Dong, and C. Change Loy, "Recovering realistic texture in image super-resolution by deep spatial feature transform," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 606–615.

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

[24] Y. Li, X. Hou, C. Koch, J. M. Rehg, and A. L. Yuille, "The secrets of salient object segmentation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 280–287.

[25] D. Martin, C. Fowlkes, D. Tal, and J. Malik, "A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics," in *Proc. 8th IEEE Int. Conf. Comput. Vis. (ICCV)*, 2001, pp. 416–423.

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

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

[28] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, "Multi-scale continuous CRFs as sequential deep networks for monocular depth estimation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 5354–5362.

[29] J. Zhang, Y. Dai, F. Porikli, and M. He, "Multi-scale salient object detection with pyramid spatial pooling," in *Proc. Asia–Pacific Signal Inf. Process. Assoc. Annu. Summit Conf. (APSIPA ASC)*, Dec. 2017, pp. 1286–1291.

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

[31] T. Wang *et al.*, "Detect globally, refine locally: A novel approach to saliency detection," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 3127–3135.

[32] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–14.

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

[34] S. Chen, X. Tan, B. Wang, and X. Hu, "Reverse attention for salient object detection," in *Proc. IEEE Eur. Conf. Comput. Vis.*, Sep. 2018, pp. 234–250.

[35] T. Wang, A. Borji, L. Zhang, P. Zhang, and H. Lu, "A stagewise refinement model for detecting salient objects in images," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 4019–4028.

[36] X. Wang, H. Ma, and X. Chen, "Salient object detection via fast R-CNN and low-level cues," in *Proc. IEEE Int. Conf. Image Process. (ICIP)*, Sep. 2016, pp. 1042–1046.

[37] X. Wang, H. Ma, X. Chen, and S. You, "Edge preserving and multi-scale contextual neural network for salient object detection," *IEEE Trans. Image Process.*, vol. 27, no. 1, pp. 121–134, Jan. 2018.

[38] A. D. Sappa and F. Dornaika, "An edge-based approach to motion detection," in *Proc. Comput. Sci. Int. Conf., Reading*, 2006, p. 563–570.

[39] J. Zhang, Y. Dai, F. Porikli, and M. He, "Deep edge-aware saliency detection," Aug. 2017, *arXiv:1708.04366*. [Online]. Available: https://arxiv.org/abs/1708.04366

[40] J. Liu, Q. Hou, M. Cheng, J. Feng, and J. Jiang, "A simple pooling-based design for real-time salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 3917–3926.

[41] W. Guan, T. Wang, J. Qi, L. Zhang, and H. Lu, "Edge-aware convolution neural network based salient object detection," *IEEE Signal Process. Lett.*, vol. 26, no. 1, pp. 114–118, Jan. 2019.

[42] Z. Deng *et al.*, "R3net: Recurrent residual refinement network for saliency detection," in *Proc. IEEE Int. Joint Conf. Artif. Intell.*, Jul. 2018, pp. 684–690.

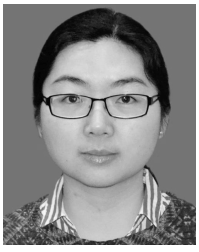
[43] P. O. Pinheiro, T.-Y. Lin, R. Collobert, and P. Dollár, "Learning to refine object segments," in *Proc. IEEE Conf. Eur. Conf. Comput. Vis.*, Sep. 2016, pp. 75–91.

[44] D. Mumford and J. Shah, "Optimal approximations by piecewise smooth functions and associated variational problems," *Commun. Pure Appl. Math.*, vol. 42, no. 5, pp. 577–685, Jul. 1989.

[45] F. Milletari, N. Navab, and S.-A. Ahmadi, "V-Net: Fully convolutional neural networks for volumetric medical image segmentation," in *Proc. 4th Int. Conf. 3D Vis. (3DV)*, Oct. 2016, pp. 565–571.

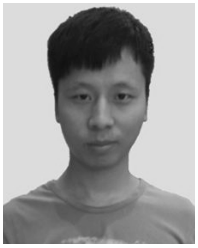
[46] A. A. Taha and A. Hanbury, "Metrics for evaluating 3D medical image segmentation: Analysis, selection, and tool," *BMC Med. Imag.*, vol. 15, no. 1, p. 29, Dec. 2015.

[47] Y. Xu *et al.*, "Structured modeling of joint deep feature and prediction refinement for salient object detection," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 3789–3798.
[48] N. Liu and J. Han, "DHSNet: Deep hierarchical saliency network for salient object detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2016, pp. 678–686.
[49] X. Li *et al.*, "DeepSaliency: Multi-task deep neural network model for salient object detection," *IEEE Trans. Image Process.*, vol. 25, no. 8, pp. 3919–3930, Aug. 2016.
[50] L. Wang *et al.*, "Learning to detect salient objects with image-level supervision," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jul. 2017, pp. 136–145.
[51] P. Zhang, D. Wang, H. Lu, H. Wang, and B. Yin, "Learning uncertain convolutional features for accurate saliency detection," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2017, pp. 212–221.
[52] W. Zhu, S. Liang, Y. Wei, and J. Sun, "Saliency optimization from robust background detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2014, pp. 2814–2821.
[53] H. Jiang, J. Wang, Z. Yuan, Y. Wu, N. Zheng, and S. Li, "Salient object detection: A discriminative regional feature integration approach," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Portland, OR, USA, Jun. 2013, pp. 2083–2090.
[54] T. Park, M.-Y. Liu, T.-C. Wang, and J.-Y. Zhu, "Semantic image synthesis with spatially-adaptive normalization," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 2337–2346.
[55] M. Everingham and J. Winn, "The PASCAL visual object classes challenge 2007 (VOC2007) development kit," *Int. J. Comput. Vis.*, vol. 111, no. 1, pp. 98–136, 2006.
[56] M. Abadi *et al.*, "TensorFlow: Large-scale machine learning on heterogeneous distributed systems," Aug. 2016, *arXiv:1603.04467*. [Online]. Available: https://arxiv.org/abs/1603.04467
[57] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *Proc. 3rd Int. Conf. Learn. Represent.*, 2015, pp. 1–15.
[58] T. Liu *et al.*, "Learning to detect a salient object," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 2, pp. 353–367, Feb. 2011.
[59] A. Borji, D. N. Sihite, and L. Itti, "Salient object detection: A benchmark," in *Proc. IEEE Conf. Eur. Conf. Comput. Vis.*, Oct. 2012, pp. 414–429.
[60] R. Achanta, S. Hemami, F. Estrada, and S. Susstrunk, "Frequency-tuned salient region detection," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2009, pp. 1597–1604.
[61] C. Li, W. Xia, Y. Yan, B. Luo, and J. Tang, "Segmenting objects in day and night: Edge-conditioned CNN for thermal image semantic segmentation," Jul. 2019, *arXiv:1907.10303*. [Online]. Available: https://arxiv.org/abs/1907.10303
[62] C. Li, X. Liang, Y. Lu, N. Zhao, and J. Tang, "RGB-T object tracking: Benchmark and baseline," *Pattern Recognit.*, vol. 96, Dec. 2019, Art. no. 106977.



Zhengzheng Tu received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2007 and 2015, respectively. She is currently an Associate Professor with the School of Computer Science and Technology, Anhui University. Her current research interests include computer vision, pattern recognition, and deep learning.

Yan Ma received the B.S. degree from Fuyang Normal University, Fuyang, China, in 2018. She is currently pursuing the M.S. degree with Anhui University, Hefei, China. Her current research is RGBT salient object detection based on deep learning.



Chenglong Li received the M.S. and Ph.D. degrees from the School of Computer Science and Technology, Anhui University, Hefei, China, in 2013 and 2016, respectively. From 2014 to 2015, he worked as a Visiting Student with the School of Data and Computer Science, Sun Yat-sen University, Guangzhou, China. He was a Post-Doctoral Research Fellow with the National Laboratory of Pattern Recognition (NLPR), Center for Research on Intelligent Perception and Computing (CRIPAC), Institute of Automation, Chinese Academy of Sciences (CASIA), China. He is currently an Associate Professor with the School of Computer Science and Technology, Anhui University. His research interests include computer vision and deep learning. He was a recipient of the ACM Hefei Doctoral Dissertation Award in 2016.

Jin Tang received the B.Eng. degree in automation and the Ph.D. degree in computer science from Anhui University, Hefei, China, in 1999 and 2007, respectively.

He is currently a Professor with the School of Computer Science and Technology, Anhui University. His current research interests include computer vision, pattern recognition, machine learning, and deep learning.

Bin Luo received the B.Eng. degree in electronics and the M.Eng. degree in computer science from Anhui University, Hefei, China, in 1984 and 1991, respectively, and the Ph.D. degree in computer science from the University of York, York, U.K., in 2002.

He is currently a Professor with Anhui University. He has authored over 200 articles in journals, edited books, and refereed conferences. His current research interests include random graph-based pattern recognition, image and graph matching, graph spectral analysis, and video analysis. He is also the Chair of the IEEE Hefei Subsection. He has served as a Peer Reviewer for international academic journals, such as the IEEE TRANSACTIONS ON PATTERN ANALYSIS AND MACHINE INTELLIGENCE, *Pattern Recognition*, *Pattern Recognition Letters*, the *International Journal of Pattern Recognition and Artificial Intelligence*, *Knowledge and Information Systems*, and *Neurocomputing*.