EDNet: Encoder-Decoder Network with Low Rank and Margin Regularizations for Saliency Detection

Jianbo Tang 1, Dongjing Shan 1, *, Weijun Li 2 and Bo Zhao 1

1 Army Engineering University Communications Officer School, Chongqing 400035, China
2 78118 troops, Chengdu 610000, China

*Corresponding author Email: shandongjing@pku.edu.cn

Abstract. Deep learning based methods have achieved great progress in the saliency detection domain. However, it remains to be challenging to segment objects in complicated scenarios consistently and accurately. To address the problem, we propose an encoder-decoder based framework in this paper, where the encoder extracts deep features for each super-pixel via fast R-CNN and attention mechanism, and a loss term is introduced to penalize the segmentation inaccuracy; the decoder makes saliency predictions by two kinds of bidirectional long short-term memory networks: the segmented BLSTM collects information horizontally and the skipping BLSTM fuse them along the vertical direction. For further improving the detection accuracy and robustness, low rank constraint is imposed on the background super-pixels, and the classification margin between salient and non-salient regions is expanded simultaneously. In experiments, we evaluate the networks on five benchmark datasets, and the results under different metrics demonstrate that our model without any post-processing has superiority over other state-of-the-art ones.

Keywords: Saliency detection, recurrent neural network, long short-term memory, low rank, super-pixel segmentation.

1. Introduction

Human visual system is able to find objects of interest in complex scenes, and saliency detection aims at highlighting the most visually obvious regions by simulating this system. The detection methods have been developed from traditional approaches to data-driven deep learning approaches in the past decades, and have been applied as a pre-processing step in many computer vision fields, such as object tracking, image editing, image retrieval, image segmentation and so on. The traditional approaches mostly exploit low-level cues and have limited ability in detection owing to lack of semantic information, the deep learning ones can model high-level semantics and have achieved promising results, but still encounter challenges when facing complicated scenarios: 1) as the loss of necessary detailed information in repeated sub-sampling of convolutional layers, deep learning models have difficulty in maintaining spatial coherence of predictions; 2) the segmentation of salient and non-salient regions is not accurate when they have similar appearances.
To address the problems, we propose a encoder-decoder model (named as EDNet) in this paper, which treats saliency detection as a task of classification on segmented super-pixels. The encoder extracts features from super-pixels by using Fast R-CNN and attention mechanism, and the decoder applies two layers of Bidirectional Long Short-Term Memory (BLSTM) to collect spatial information, i.e., the segmented BLSTM absorbs information horizontally along the rows, and the skipping BLSTM fuses information across them. Bidirectional mode allows the cells in two ends of the time-line convey the same amount of information. Moreover, as the image background (non-salient regions) usually lies in a low-dimensional subspace and the salient regions deviate from this subspace, a low rank regularization is utilized to penalize the background outputs of the skipping BLSTM during training, and then the classification margin is expanded via a special loss function, the two regularization is beneficial for improving the detection accuracy in complex scenes.

The main contributions can be summarized as three folds:

● We propose an Encoder-Decoder framework for saliency detection, whose encoder and decoder are based on super-pixel segmentation and new designed BLSTMs respectively. The framework can guarantee spatial coherence and sharp contours owing to integration of super-pixels in deep learning pipeline.

● We propose to impose low rank constraint on the background super-pixels and expand the classification margin between salient and non-salient regions, which can make the saliency prediction more robust.

● We evaluate the proposed methods with two kinds of backbones on five benchmark datasets, and they have achieved competitive performance compared with 17 previous state-of-the-art methods under different metrics.

2. Related Work

2.1. Traditional models

Traditional models usually extract hand-crafted features from super-pixels and are mostly based on low-level cues. For instance, early classic algorithms are typically based on the priors of regional center-surround contrast or global contrast; and later on, the methods exploiting intrinsic property of saliency are researched, which are built on various mathematical models, such as low rank matrix recovery [1, 2], principal component analysis (PCA), graphical model or manifold ranking. Among them, the work in [1] introduced an overcomplete dictionary to represent image patches, and estimated saliency information via low-rank and sparsity matrix decomposition; the work in [2] proposed to distinguish background from foreground by low rank property of the background’s appearances, and located salient regions by direction of sparse noises that are generated via Robust PCA. Because of the lack of semantic information or high-level cues, these traditional methods have limited ability to detect salient objects in complex scenes. Thus, deep learning models emerge as the times require.

2.2. Deep learning models

In recent years, a large number of deep learning approaches emerge owing to the improvement in abundant data processing capability, which can model high-level semantics and have achieved promising results in saliency detection. The methods based on Fully Convolutional Network (FCN), such as [3], have been widely used in this field, which infer saliency from global perspective. Nonetheless, the methods are prone to miss edge information of salient objects, and also need to enhance the spatial coherence of their predictions. Some methods cope with the problems by using a combination or refinement method, for instance, Peng et al. [4] proposed to combine the local and global saliency cues and use composition information to obtain final saliency. Nevertheless, the scheme may cause computational redundancy and training difficulty, which is not conducive to successive tasks and real applications. Our proposed model combines convolutional feature map and super-pixel segmentation, and perform predictions on the super-pixels by BLSTM networks, which is different from the previous models and does not need refinement operations.
3. Proposed Method

In the following parts, the components of the proposed neural network, i.e., an encoder and a decoder (see Fig.1), are illustrated subsequently, and then the low rank and margin regularizations used in the training strategy are also provided.

3.1. Feature Extraction by Encoder

Fast R-CNN is usually utilized in object detection, where an image and multiple candidate bounding boxes (i.e., regions of interest: Rois) are input into a fully convolutional network, and then each bounding box is pooled into fixed-size feature map, and finally fixed-size representation is generated by fully 90 connected layers (FCs).

![Framework of the EDNet](image)

**Fig. 1** Framework of the EDNet

The Encoder module aims at extracting features for each super-pixel and the Decoder module is responsible for predicting saliency values, which employs low rank and margin regularizations for improving classification accuracy.

In this paper, we apply the Fast R-CNN to perform feature extraction for saliency detection, in which the super-pixel segmentation can help to preserve important image properties like boundaries and
eliminate hollow effect in salient maps. The complete process of feature extraction is depicted in the lower part of Fig.1: at first, a feature map is computed on an image by the ResNet network that comprises 34 convolutional layers, and simultaneously, the image is segmented by using SLIC method; secondly, as the original size of the feature map is small, a deconvolution layer is used to expand it and then facilitate the Roi projection, which can project the bounding box of each super-pixel to the 100 feature map; next, the cropped features are pooled into fixed size by Roi pooling layer; finally, dimension reduced features are generated by a fully connected layer (FC), and rearranged into a long sequence with fixed-length entries.

Subsequently, we introduce an attention mechanism for the feature of each super-pixel. In some cases, the non-salient regions may have "salient-like" appearance, or the salient regions may look different from the main body and be prone to be ignored in detection. Human vision can avoid the mistakes by examining the surroundings carefully, then it is reasonable for neural networks to simulate the way by using attention mechanism, which can expand the perception field for each super-pixel and alleviate the confusions. The typical attention mechanism is very time-consuming as its perception field spreads all over the time line. We propose a self-attention model on 1-ring neighbors in this paper, i.e., the perception field of each super-pixel is constrained in its surrounding area. The output of the attention layer $a_i$ is calculated by linear combination on the regional features:

$$ a_i = \sum_{j \in N(i)} \lambda_{ij} f_j $$

(1)

Where $N(i)$ is the 1-ring neighbors of the i-th super-pixel, $f_j$ represents the feature vector of the j-th super-pixel, and $\lambda_{ij}$ is the combination coefficient, which is calculated by:

$$ \lambda_{ij} = \frac{e^{\text{Sim}_{ij}}}{\sum_{j \in N(i)} e^{\text{Sim}_{ij}}} , \quad \text{Sim}_{ij} = f_i ^T M f_j, $$

(2)

Where $M$ is weight matrix of a fully connected layer, Sim$_{ij}$ represents the similarity score between the two feature vectors. This mechanism can make the detection of the object details more robust, and the computation efficiency also can be maintained as the attention scope of each super-pixel is limited in its adjacent neighbors, which results in that the times of similarity comparison is reduced sharply compared to the conventional global mode.

3.2. Saliency Prediction by Decoder

3.2.1. Structures of Bidirectional LSTMs. As the structures of our designed Bidirectional LSTMs are based on the standard LSTM, its cell structure is presented below as a pre-requirement for understanding.

$$ f_i = \sigma (W_f h_i + U_f y_{i-1} + b_f) $$
$$ i_i = \sigma (W_i h_i + U_i y_{i-1} + b_i) $$
$$ g_i = \tanh (W_g h_i + U_g y_{i-1} + b_g) $$
$$ o_i = \sigma (W_o h_i + U_o y_{i-1} + b_o) $$
$$ c_i = f_i \odot c_{i-1} + i \odot g_i $$
$$ h_i = o_i \odot \tanh (c_i) $$

(3)

where $\{f, i, o, g\}$ symbolize the forget, input, output gates and data processing respectively, in which $\sigma$ is short for sigmoid function; $c_i$ and $h_i$ represent the cell and hidden states.

In order to collect the spatial information among super-pixels, two kinds of Bidirectional LSTM structures are proposed and utilized here: the segmented and skipping BLSTMs. In bidirectional mode
of LSTMs, a series of super-pixel features are fed as inputs with different directions, which allows the network to capture the dependency relations among super-pixels by following two sequential orders. In line by line scanning mode, super-pixels are interrupted by image boundaries, thus we proposed a segmented LSTM and use it in the first layer of the decoder, its structure is formulated as:

\[
\begin{align*}
    f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
    i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
    s_t &= f_{\text{binary}}(\sigma(W_s x_t + U_s h_{t-1} + b_s)) \\
    g_t &= \tanh(W_g x_t + U_g h_{t-1} + b_g) \\
    o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
    c_t &= f_t \odot c_{t-1} \odot (1 - s_{t-1}) + i_t \odot g_t \\
    h_t &= o_t \odot \tanh(c_t)
\end{align*}
\]  

where the symbols have the same meaning as those in the standard LSTM, the difference lies in that we introduce a segmented gate \( s_t \) to truncate the state flow at intervals of the image boundaries, where \( f_{\text{binary}}(\cdot) \) is a deterministic step function that outputs binary values, and we use a straight-through estimator to compute the gradients during training. Note that \( s_t \) is a scalar and should be repeated as vector \( \bar{s}_t \) when used in dot products.

Next, another LSTM layer should be used to collect information across the scanning lines. In order to satisfy the low rank property illustrated in Section 3.2.2, we propose a skipping LSTM that can skip state updates when encountering super-pixels with similar BN appearance, its structure is defined as:

\[
\begin{align*}
    f_t &= \sigma(W_f x_t + U_f h_{t-1} + b_f) \\
    i_t &= \sigma(W_i x_t + U_i h_{t-1} + b_i) \\
    p_t &= f_{\text{binary}}(\sigma(W_p x_t + U_p h_{t-1} + b_p)) \\
    g_t &= \tanh(W_g x_t + U_g h_{t-1} + b_g) \\
    o_t &= \sigma(W_o x_t + U_o h_{t-1} + b_o) \\
    c_t &= (1 - p_t) \odot (f_t \odot c_{t-1} + i_t \odot g_t) + p_t \odot c_{t-1} \\
    h_t &= o_t \odot \tanh(c_t)
\end{align*}
\]  

Where \( p_t \) is a skipping gate and used to determine whether the state should be updated or copied from the previous timestamp. Apparently, LSTM with copied states are more likely to generate low rank output matrix, and in turn a low rank regularization can drive the LSTM to filter out perturbations in similar super-pixels and skip relevant state updates.

### 3.2.2 Classification with Low Rank and Margin Regularizations.

Saliency detection is handled as a binary classification problem (salient or non-salient) in this method. As depicted in the framework of Fig.1, the outputs from the skipping LSTM are passed through a full connected layer for dimension reduction and a Softmax function for category labels, and then cross entropy is adopted as loss function. In order to improve the accuracy of classification, we design two regularizers based on the principle of minimum-in-cluster-distance and maximum-between-cluster-distance. Given a sequence of ground truth \( G = \{g_1, \ldots, g_n\} \) and final outputs of the network \( Y = \{y_1, \ldots, y_n\} \), the objective function can be formulated as:

\[
\begin{align*}
    &\min_{\theta} l(G, Y, \theta), \\
    &\text{s.t.} \text{rank}(H_x) < r
\end{align*}
\]
and \(\min_{\substack{\theta \in \Theta, h_i, h_j \in H \setminus \emptyset \geq 0}} ||h_i - h_j||_2^2 \geq \text{margin}\)

Where \(\theta\) represents a set of the network parameters, and \(L\) indicates a cross entropy loss function. Outputs of the skipping LSTM is denoted as \(H = \{h_1, ..., h_n\} \in \mathbb{R}^{D \times H}\) where \(D\) is the number of hidden units in the LSTM, \(N\) is the number of super-pixels, then we use a label matrix \(Q\) to indicate.

Whether or not the entry \(h_i\) belongs to salient regions, we have \(Q = [q_1, ..., q_N] = I - \text{diag}(g_1, ..., g_N)\), where \(I\) is an identity matrix and \(g_i\) is the ground truth label. If \(g_i = 0\), then \(q_i = 1\) which means that \(h_i\) represents non-salient region and vice versa. Next, we can get a matrix that only comprises information corresponding to background super-pixels (i.e., non-salient regions): \(H_s = HQ\), where \(h_i\) related to salient super-pixels become zero-vector as they are multiplied by \(g_i\). Owing to homogeneity of the image background, low rank constraint can be imposed on the matrix \(H_s\) (\(r\) is small scalar set to 16 in experiments), and the skipping LSTM can help to avoid linear proportion relations among the entries (e.g., \(h_i = 2h_j\) can keep low rank but may fall into two categories).

\[\begin{array}{c}
Y \\
L_2 \\
\theta_{fc} \\
L_1 \\
H_n \\
\theta_{l} - r = \theta_{l} - \theta_{r} \\
F \\
0_r \\
\text{Img}
\end{array}\]

**Fig. 2** Forward propagation of the neural network

In figure 2: \(\text{Img}\) represents an input image; \(\theta_{r} - \theta_{r}\) indicates a parameter set that excludes \(\theta_{l}\) from \(\theta_{l}\).

Thus, the inner distance of non-salient super-pixels can be reduced by this regularizer, and a more robust classifier can be obtained as representations lying in a low-dimensional subspace are robust to input perturbations. Similarly, \(H_s = H(I - Q)\) only has information corresponding to salient super-pixels, during training, a classification margin between representations of salient and non-salient regions can be established by the other regularizer.

As declared in [5], a low rank matrix can be projected onto itself by using a low rank projection. Then the matrix \(H_n\) can be driven to be low rank by introducing an auxiliary parameter weight matrix: \(A \in \mathbb{R}^{D \times D}\). The objective function is advanced to:

\[
\min_\theta l(G, Y, \theta) + L_s(H_s; A, \theta) + L_s(H_s, H_s; \theta),
\]

\[s.t. \text{rank}(A) \ll D, \quad (7)\]

Where \(\theta_{l}\) represents the parameters in part of the network under the outputs \(H\) in a forward pass, and

\[
L_s(H_s; A; \theta) = ||AH_s - H_s||_2^2
\]

\[s.t. \text{rank}(A) \ll D, \quad (7)\]
\[ L_2(H_x;H_y;\theta) = \max \left\{ \text{arg min}_{h_i,h_j} \| h_i - h_j \|_2^2, 0 \right\}, \quad \text{s.t. } h_i \in H_x, h_j \in H_y; h_i, h_j \neq 0, \]  

(8)

Algorithm 1 Optimization scheme;
Input: Input images, learning rate \( \eta \);
Parameter: Network weights \( \theta \), auxiliary matrix \( A \), and \( W_d \);
Output: \( \theta^{*} \);
1: Initialize \( \theta^{(0)} - \theta^{(r)} \), \( A \) and \( W_d \) by Xavier uniform initializer; load pretrained ResNet-50 \( \theta^{(r)} \).
2: while \( e \leq \) epoch number do
3: while \( i \leq \) batch number do
4: \[ \nabla_{\theta^{(r)}} = \frac{\partial L_2}{\partial \theta^{(r)}}; \nabla_A = \frac{\partial L_2}{\partial A} \]
5: \[ \nabla_{\theta^{(r)}} = \frac{\partial L_1 + L_2}{\partial \theta^{(r)}} \]
6: \[ \nabla_{\theta^{(r)}} = \frac{\partial L_1 + L_2}{\partial W_d} \]
7: // {Update parameters by gradient descent}
8: \( (\theta, A) = (\theta^{(r)}, A) - \eta \nabla_{(\theta, A)} \)
9: // {Hard thresholds the rank of A by using low rank matrix approximation}
10: \( A \leftarrow \Pi^{\text{rank}}_{\theta^{(r)}}(A) \)
11: end while
12: end while
13: return \( \theta^{*} \) with the minimum loss on validation set.

In the \( L_1 \) term of equation (8), if \( A \) is low rank, then \( A = AH_n \) is low rank as \( \text{Rank}(AH_n) \leq \min \{ \text{rank}(A), \text{rank}(H_n) \} \). During training, \( A \) is made to be low rank by low rank matrix approximation based on singular value decomposition (SVD), the scalar \( r \) is used as a hard threshold to preserve large singular values and eliminate others. In the \( L_2 \) term, margin is set to 1 in experiments.

The forward pass process of the neural network is presented in Fig.2, where the circles represent intermediate results output by the inner network layers, and the subsets of parameters are attached along the gradient propagation path. The optimization scheme is summarized in Algorithm 1.

4. Experiments

4.1. Datasets and evaluation metrics

The proposed model is evaluated on five public saliency detection datasets. DUTS is comprised of 10,553 images in the training set (DUTS-TR) and 5,019 images in the test set (DUTS-TE); ECSSD [6] contains 1,000 complex images with multiple objects of different sizes; DUT-OMRON [6] consists of 5,168 challenging images, each of which has one or two salient objects in cluttered background; PASCAL-S [6] contains 850 well annotated natural images; HKU-IS [5] contains 4,447 challenging images with low color contrast or multiple salient objects.

For comparison, we adopt four widely used metrics: PR curves (precision-recall), F-measure, E-measure, S-measure and MAE (mean absolute error). PR curves are used to evaluate the models performance quantitatively and can be plot based on the (Precision, Recall) pairs, to calculate these pairs, binary masks are generated from a saliency map by sliding a threshold from 0 to 255, and then compared against the ground truth mask. F-measure, Fm considers the precision and recall pairs and
computes a F-score for each saliency map, and then the scores are averaged on all the images. E-measure, Em evaluates the similarity by jointly capturing image-level statistics and local pixel matching information. S-measure, Sm evaluates the object-aware and region-aware structure similarities between the prediction and the ground truth. MAE measures the difference between the saliency map and the ground truth mask in pixel level.

4.2. Implementation details

The input image in our experiment are resized into 320 × 320 pixels for testing, and 160 super-pixels are adopted for segmentation. In the encoder module, the backbone model ResNet-50 is initialized with ImageNet weights, then its generated feature maps are expanded to 64 × 64 by using a pooling layer and sequentially to 128 × 128 by a deconvolution layer with stride of 2, next, the bounding boxes of super-pixels are projected onto the feature maps by Roi-projection, and through 7 × 7 Roi pooling, 512×7×7-dimensional features are extracted for each super-pixel, which are then compressed to 512-dimensional vectors by a FC layer. Afterwards, in the decoder module, the feature vectors are input into Bidirectional LSTM layers as sequential data, the numbers of hidden units in the two bidirectional LSTM layers are set to 1024 (segmented LSTM) and 512 (skipping LSTM) respectively. Finally, the outputs of the top layer are used for classification via cross entropy loss function.

Our network is trained on DUT-TR dataset with a mini-batch of 16 by using two Adam optimizers for the encoder and decoder. The learning rates are both set to 0.001 and decayed by half when the validation loss does not descend for 10 epoches. The MSRA1000 that comprises 1000 images selected randomly from MSRA10k is adopted as the validation set, the whole training process lasts for 300 epoches, and the model parameters with the best validation accuracy are saved as the trained model. Additionally, we test our network with the backbone of VGG-16 network, all the hyperparameters are kept the same with those used for the backbone of ResNet-50.

4.3. Comparative results

We compare our models (EDNet-V and EDNet-R using VGG-16 and ResNet-50 respectively) with 17 state-of-the-art models, namely Amulet17 [7], NLDF17 [8], DSS17 [3], MLMS19 [9], AFNet19 [10], LFR19 [11], GateNet-VA [6], PiCANet18 [12], DGRL18 [13], CapSal [14], BASNet [15], CPD19 [16], and GateNet-R20 [6], where the sub-scripts indicate the year of publication, and all saliency maps of these models are provided by authors.

The comparative results in terms of Fm, Em, Sm and MAE score on five benchmark datasets are presented in Table 1, and the approaches in the top part of the table take backbones of VGG-16, others in the bottom part take backbones of ResNet-50 unless especially stated. The values ranking first, second and third are highlighted in Red, Green and Blue respectively. As seen from the table, our models achieve the best performance in most cases under the same backbones and behave better than any of others on the whole. Moreover, we exhibit the standard PR curves and the F-measure curves in Fig.3. Our model with backbone of ResNet-50 (EDNet-R in black solid line) achieves the best results on the DUTS-TE, DUT-OMRON, PASCAL-S and HKU-IS datasets, and behaves competitively on the ECSSD. Also we can see that EDNet-R improves a lot than EDNet-V that uses backbone of VGG-16.

Some resulted saliency maps are selected and presented in Fig.4, which are representative for the challenging cases, such as complicated backgrounds, low contrast of the appearance with its surroundings, and multiple salient objects. Owing to the super-pixel segmentation module adopted in the encoder of our framework, the detected salient objects have more sharp and clear contours than those of other approaches; self-attention mechanism in the encoder can associate every super-pixel with its surroundings and help to recognize salient-like non-salient regions or salient regions with low contrast appearance compared to the background, thus our model can detect more details and the saliency maps are closer to the ground truth, e.g., in Fig.4, the detected jet of our models is more complete than that of others; as the saliency value for each super-pixel is assigned with either 1 or 0, the hollow effect inside the salient objects can be restrained to a great extent, there exist some hollows or cracks on the saliency
maps of DSS and RAS, while our model can eliminate the phenomena; besides, our model show good potential in detecting multiple objects, such as in the third image of Fig.4, EDNet-R can detect the three persons accurately.

Table 1. Quantitative evaluation results

| Dataset | DUTS-TE | ECSSD | DUT-OMRON | PASCAL-S | HKU-IS |
|---------|---------|-------|------------|----------|--------|
|         | Fm, Em, Sm, MAE | Fm, Em, Sm, MAE | Fm, Em, Sm, MAE | Fm, Em, Sm, MAE | Fm, Em, Sm, MAE |
| ANN[7]  | 0.678, 0.803, 0.804, 0.803 | 0.866, 0.912, 0.894, 0.809 | 0.647, 0.784, 0.780, 0.658 | 0.771, 0.831, 0.819, 0.808 | 0.843, 0.915, 0.889, 0.809 |
| NLDF[12] | 0.739, 0.855, 0.855, 0.856 | 0.878, 0.912, 0.875, 0.865 | 0.684, 0.817, 0.770, 0.680 | 0.782, 0.842, 0.864, 0.814 | 0.873, 0.929, 0.878, 0.848 |
| DSS[13] | 0.716, 0.843, 0.843, 0.844 | 0.873, 0.915, 0.873, 0.863 | 0.674, 0.810, 0.790, 0.674 | 0.770, 0.830, 0.795, 0.710 | 0.850, 0.920, 0.878, 0.750 |
| MLMS[14] | 0.745, 0.863, 0.863, 0.864 | 0.868, 0.916, 0.911, 0.944 | 0.692, 0.829, 0.868, 0.664 | 0.771, 0.847, 0.844, 0.705 | 0.871, 0.938, 0.906, 0.809 |
| AFNet[15] | 0.793, 0.895, 0.895, 0.895 | 0.906, 0.941, 0.914, 0.942 | 0.738, 0.859, 0.826, 0.657 | 0.828, 0.887, 0.850, 0.707 | 0.888, 0.947, 0.906, 0.806 |
| LFRL[16] | 0.742, 0.853, 0.853, 0.853 | 0.911, 0.946, 0.914, 0.942 | 0.718, 0.851, 0.811, 0.664 | 0.813, 0.851, 0.851, 0.703 | 0.860, 0.914, 0.914, 0.806 |
| GateNet+V[8] | 0.783, 0.888, 0.888, 0.888 | 0.896, 0.931, 0.917, 0.941 | 0.723, 0.852, 0.826, 0.661 | 0.803, 0.894, 0.790, 0.707 | 0.889, 0.947, 0.906, 0.806 |
| EDNet-V[9] | 0.806, 0.902, 0.871, 0.871 | 0.912, 0.947, 0.925, 0.934 | 0.727, 0.859, 0.821, 0.656 | 0.824, 0.861, 0.827, 0.707 | 0.901, 0.952, 0.909, 0.901 |

| Backbone: VGG-16 | |
|------------------|------------------|------------------|------------------|------------------|------------------|
| PiCANet[12]      | 0.759, 0.873, 0.869, 0.854 | 0.886, 0.927, 0.917, 0.946 | 0.717, 0.848, 0.832, 0.665 | 0.804, 0.862, 0.854, 0.708 | 0.870, 0.940, 0.905, 0.843 |
| DGBL[13]         | 0.794, 0.890, 0.890, 0.890 | 0.906, 0.946, 0.906, 0.941 | 0.733, 0.856, 0.810, 0.662 | 0.827, 0.901, 0.850, 0.703 | 0.909, 0.940, 0.939, 0.806 |
| CapsNet[14]      | 0.755, 0.807, 0.818, 0.803 | 0.825, 0.866, 0.826, 0.704 | 0.564, 0.700, 0.673, 0.696 | 0.627, 0.728, 0.827, 0.704 | 0.641, 0.905, 0.934, 0.806 |
| BASNet[15]       | 0.791, 0.884, 0.884, 0.884 | 0.888, 0.921, 0.916, 0.937 | 0.756, 0.866, 0.866, 0.656 | 0.791, 0.853, 0.821, 0.707 | 0.905, 0.924, 0.918, 0.802 |
| CFPE[16]         | 0.805, 0.904, 0.898, 0.894 | 0.917, 0.949, 0.918, 0.937 | 0.747, 0.873, 0.824, 0.656 | 0.831, 0.887, 0.844, 0.702 | 0.891, 0.950, 0.906, 0.804 |
| GateNet+R[8]     | 0.807, 0.903, 0.884, 0.884 | 0.916, 0.943, 0.920, 0.940 | 0.716, 0.868, 0.837, 0.655 | 0.819, 0.855, 0.792, 0.669 | 0.899, 0.953, 0.915, 0.833 |
| EDNet-R          | 0.828, 0.910, 0.883, 0.888 | 0.931, 0.950, 0.920, 0.922 | 0.766, 0.871, 0.811, 0.655 | 0.852, 0.902, 0.853, 0.705 | 0.922, 0.962, 0.922, 0.827 |

* In GateNet, "-V" means using VGG-16 as backbone, and "-R" means using ResNet-50;

* In CapsSal, the backbone of ResNet-101 is utilized;

* In BASNet, the backbone of ResNet-34 is utilized.

Fig. 3 PR and F-measure curves on five benchmarks.
Fig. 4 Qualitative comparison of the models

(Including the EDNet-V with backbone of VGG-16 and the EDNet - R with backbone of ResNet-50). Testing images we choose are challenging for detection, like low contrast (the first image), complicated background (the second image), multiple ad small objects (the third and fourth images. Besides, ‘GT’ represents the ground truth.)

5. Conclusions
In this paper, we propose an Encoder-Decoder based framework for saliency detection. At first, in the encoder module, features are extracted for each super-pixel by using Fast R-CNN, and then, the receptive fields of super-pixels are extended to their 1-ring neighbors by attention mechanism, lastly in the decoder module, segmented and skipping BLSTMs are used for exploiting spatial information and predicting saliency values on the super-pixels. During network training, the proposed low rank and margin regularizations are utilized to make the classification (salient or non-salient w.r.t each super-pixel) more accurate. We conducted experiments on 5 benchmark datasets and adopted different kinds of metrics to evaluate the performance, numerous previous state-of-the-art approaches are applied for comparison, and the quantitative results like PR-curves and qualitative saliency maps are proofs of the superior performance of the model we proposed.

References
[1] J. Yan, M. Zhu, H. Liu, Y. Liu, Visual Saliency Detection via Sparsity Pursuit, IEEE Signal Processing Letters 17 (2010) 739-742. doi: 10.1109/LSP.2010.2053200.
[2] X. Shen, Y. Wu, A unified approach to salient object detection via low rank matrix recovery, in: Computer Vision and Pattern Recognition, 2012, pp. 853-860.
[3] Q. Hou, M.-M. Cheng, X. Hu, A. Borji, Z. Tu, P. Torr, Deeply supervised salient object detection with short connections, IEEE TPAMI 41 (4) (2019) 815-828.
[4] Q. Peng, Y. M. Cheung, X. You, Y. Y. Tang, A hybrid of local and global saliencies for detecting image salient region and appearance, IEEE Transactions on Systems Man & Cybernetics Systems (2016) 1-12.
[5] A. Sanyal, V. Kanade, P. H. S. Torr, Low rank structure of learned representations, CoRR abs/1804.07090. arXiv: 1804.07090.
[6] X. Zhao, Y. Pang, L. Zhang, H. Lu, L. Zhang, Suppress and balance: A simple gated network for salient object detection, 2020.
[7] P. Zhang, D. Wang, H. Lu, H. Wang, X. Ruan, Amulet: Aggregating multi-level convolutional features for salient object detection, CoRR abs/1708.02001. arXiv: 1708.02001. URL http://arxiv.org/abs/1708.02001.
[8] Z. Luo, A. K. Mishra, A. Achkar, J. A. Eichel, S. Li, P. Jodoin, Nonlocal deep features for salient object detection, in: 2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017, IEEE Computer Society, 2017, pp. 6593-6601. doi: 10.1109/CVPR.2017.698.
[9] R. Wu, M. Feng, W. Guan, D. Wang, H. Lu, E. Ding, A mutual learning method for salient object detec
detection with intertwined multi-supervision, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 8150-8159.

[10] M. Feng, H. Lu, E. Ding, Attentive feedback network for boundary-aware salient object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 1623-1632.

[11] P. Zhang, W. Liu, H. Lu, C. Shen, Salient object detection with lossless feature reflection and weighted structural loss, IEEE Transactions on Image Processing 28 (6) (2019) 3048-3060.

[12] N. Liu, J. Han, M. Yang, Picante: Learning pixel-wise contextual attention for saliency detection, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, IEEE Computer Society, 2018, pp. 3089-3098.

[13] T. Wang, L. Zhang, S. Wang, H. Lu, G. Yang, X. Ruan, A. Borji, Detect globally, refine locally: A novel approach to saliency detection, in: 2018 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2018, Salt Lake City, UT, USA, June 18-22, 2018, IEEE Computer Society, 2018, pp. 3127-3135.

[14] L. Zhang, J. Zhang, Z. Lin, H. Lu, Y. He, Capsal: Leveraging captioning to boost semantics for salient object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 6024-6033.

[15] X. Qin, Z. V. Zhang, C. Huang, C. Gao, M. Dehghan, M. J`agersand, Basnet: Boundary-aware salient object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 7479-7489.

[16] Z. Wu, L. Su, Q. Huang, Cascaded partial decoder for fast and accurate salient object detection, in: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, Computer Vision Foundation / IEEE, 2019, pp. 3907-3916.