Salient object detection based on Two-stream edge attention guidance

Li Wei†, Li chunlei1, Miao yue2

1 Key Laboratory of ICSP, Ministry of Education, Anhui University, Hefei 230601, China
2 Key Laboratory of ICSP, Ministry of Education, Anhui University, Hefei 230601, China
* School of Computer Science and Technology, Anhui University, Hefei 230601, China
Corresponding author’s e-mail: le998@sohu.com

Abstract: The existing U-Net structure convolutional neural network is widely used in salient object detection, but the problem of target boundary blur is caused by the convolution pooling operation. In order to keep the edge structure of salient objects clear, a two-stream edge attention guided convolutional neural network (SEANet) method is proposed to strengthen the guided learning of edge features, mainly by guiding the edge features to compare the original image features and the depth features are enhanced to improve the accuracy of salient object detection. The loss is minimized by using consistent cross entropy loss and IOU loss to maximize the coincidence rate of the true value map and the predicted map, as well as the actual boundary and the predicted boundary. At the same time, the obtained prediction map is compared with the current mainstream nine models, and good results have been achieved. Four sets of experiments are carried out in the ablation experiment, and the experimental results also confirmed that the model has a great improvement effect on the performance of salient object detection.

1. Introduction

The salient object detection is to quickly and effectively identify the salient part of one or more images containing useful information, and the detection model should at least satisfy the possibility of losing salient areas and marking the background incorrectly as salient areas. At the same time, the saliency map should have a high resolution to accurately locate the prominent object and retain the characteristic information of the original image, and the final detection model should be fast and effective. As a preprocessing part, salient object detection has been widely used in image segmentation, face recognition, visual tracking and other directions.

Traditional salient object detection mostly uses image internal clues to manually extract feature information, such as based on contrast [1], color prior region [3], center prior [4]. In recent years, with the development and improvement of convolutional neural networks in different image tasks, deep learning has gradually become the mainstream method of salient object detection. The experimental results of the existing detection model based on the fully convolutional neural network [5,6,7] also show good performance. However, frequent context switching and continuous convolution reduce the size of the feature map and deteriorate the edges of the protrusions, resulting in fuzzy results and insensitive
to the details of the image. Secondly, the classification of each pixel does not fully consider the pixel the relationship lacks spatial consistency.

Now when using convolutional neural networks for detection, people have begun to pay attention to these issues, especially boundary issues. For example [8,9], in summary, the method based on convolutional neural network still has the defect that it is difficult to maintain a clear boundary and other spatial structure.

In order to solve this problem, we propose a convolutional neural network method (SEANet) based on two-stream edge attention-guided local features for salient object detection. First, we use VGG-16 as the backbone extraction network, and each block extracts the RGB and depth features of the image. Secondly, before the detection, the initial image is preprocessed to obtain the edge feature map, and the edge attention guidance module is introduced for feature extraction and spatial transformation, which effectively realizes edge embedding. Third, considering the importance of local contrast as a significant feature of the difference between foreground and background regions, we use contrast to complement and enhance features. Finally, the local contrast feature guided by the edge attention is combined with the global context feature to generate a saliency map. On the four benchmark data sets we tested, we compared the proposed method with many of the most advanced saliency detection methods, and our methods have achieved good results.

2. Network Architecture
The Architecture model of this article is a two-stream structure, as can be seen from Figure 1, the input is the original image and the depth image, and the edge feature is introduced through the edge attention guidance module, which we will introduce in detail in section 2.2. Then through the contrast module for feature enhancement, we will introduce in detail in section 2.3. Finally, the information features of the two streams are deconvolved and up-sampled, and the final prediction image is obtained by fusing the features of the first layer. In terms of loss, we use cross-entropy loss and IoU loss to calculate the loss calculation, which we will introduce in detail in Section 3.

2.1 Overview of Network Architecture
Our SEANet selects the first 5 blocks of VGG-16 as the edge output of RGB and depth maps. Each
The convolution block uses a kernel size of 3×3 and an output feature channel of 128. The purpose of the convolution block is to obtain multiple levels.

The middle layer is where we propose the edge attention module. We use the edge map \( X^E_i \) (\( i \in \{1,2,3,4,5\} \)) obtained by preprocessing as the input of the edge guidance block. Secondly, edge features \( X^E_i \) and hierarchical multi-scale feature \( X_i \) generate edge perception features \( F^E_i \).

Finally, the edge feature \( F^E_i \) is obtained by the average pooling operation to obtain the contrast feature \( C^E_i \). The operation of the deep tributary is the same as the RGB tributary. Finally, we use deconvolution and upsampling operations to fuse two stream features, and use two convolutional layers and a softmax operation to calculate the score to get the final prediction image.

### 2.2 Edge Attention Model

The structure diagram of the traditional spatial attention module [10] is shown in Figure 2. Pass the input feature map through MaxPooling and AvgPooling, and perform feature learning and dimensionality reduction after a convolution. Finally, the activation function makes the image smooth and generates a spatial attention feature map. The calculation formula is:

\[
F^S(f) = S(c(C(\text{Max}(f)), \text{Avg}(f))))
\]

Where \( S \) is the Sigmoid activation function, \( c \) is the convolution operation, \( C \) is the connection operation, and Max and Avg are the MaxPooling and AveragePooling operations.

We introduce edge information into the traditional spatial attention mechanism for feature enhancement. The model diagram is shown in Figure 3. First, the edge feature \( X^E_i \) and the RGB feature \( f^R_i \) are multiplied by feature elements as the input of the attention module. After Pooling and Softmax operations, the edge feature \( X^E_i \) and the enhanced feature are added with elements to obtain the final edge attention perception feature. The calculation formula is:

\[
F^{Ei}(X^E_i, f^R_i) = X^E_i \odot f^R_i \oplus X^F_i
\]

Where \( \odot \) is the element multiplication operation and \( \oplus \) is the element addition operation.
2.3 Contrast
Contrast\(^{[11]}\) refers to the measurement of different brightness levels between light and dark areas in an image. The larger the difference range, the greater the contrast, the clearer the image and the more vivid colors. On the contrary, the smaller the contrast, the image is gray and black and white. The contrast is more obvious and significant.

In object detection, contrast reflects the global or local relationship between salient objects in the foreground and background. If we want to extract salient objects in the foreground, we must make the salient features evenly distributed in the foreground while preventing the background area from interfering with object detection. Obtaining contrast information is particularly important at this time.

In order to obtain contrast information, the edge attention perception feature \(F_{At}^{E_i}\) is introduced as the input of the contrast module. Each contrast feature is defined as the input feature minus its average pooling feature, where the average pooling area is \(3 \times 3\). The calculation formula is:

\[
F_{Con}^{E_i} = F_{At}^{E_i} - \text{Avg}(F_{At}^{E_i})
\]

The average pooling operation can reduce the error caused by the variance of the estimated value caused by the size limit of the neighboring area. Therefore, subtracting the average pooling feature can retain more foreground information and make it easier to detect salient objects.

2.4 Deconvolution
After comparison and calculation, the image output size is gradually reduced to \(11 \times 11\). At this time, deconvolution is used to make the output image size the same as the original image input size is \(176 \times 176\), the deconvolution block kernel size is \(5 \times 5\), and the stride is 2. In the deconvolution block of each layer, the edge attention perception feature \(F_{At}^{E_i}\) and contrast feature \(F_{Con}^{E_i}\) of the current layer and the deconvolution feature \(D_{i+1}\) of the up-sampled previous layer output are used as the input of this layer to obtain the deconvolution feature \(D_i\) of the current layer. The calculation formula is:

\[
D_i = \text{Dec}(F_{At}^{E_i}, F_{Con}^{E_i}, \text{up}(D_{i+1}))
\]

Where Dec is the deconvolution operation, and up is the up-sampling operation.

3. Loss Function
The loss function is a function that maps the value of a random event or its related random variable to a non-negative real number to represent the "risk" or "loss" of the random event. David Mumford et al.\(^{[12]}\) believe that the loss calculation of salient object detection can be attributed to the optimization of the non-convex energy function, which consists of a data term and a regular term, and this calculation method is accepted and used by most people. Based on this method, many people have improved the loss function method\(^{[13,14]}\). Although the calculation method has high accuracy, they are all based on iterative calculations, which makes them require higher initial data. When there is large noise interference, foreground background These methods are easy to fail when the difference is small and the boundary is not obvious. In order to solve these problems, Luo et al.\(^{[16]}\) proposed a combined loss function based on the sum of cross-entropy loss and IoU boundary loss. The formula is as follows:

\[
\text{Loss} = \sum_j \gamma_j (1 - \text{IoU}(E_j, \hat{E}_j)) + \sum_j \lambda_j S_j(p(v), \hat{p}(v)) dv
\]

Where \(S_j\) is the sum of the cross entropy of the ground truth map \(p(v)\) and the prediction map \(\hat{p}(v)\) of all pixels in the image area \(\Omega_j\), and \(\text{IoU}(E_j, \hat{E}_j)\) is the intersection of the pixels on the real boundary \(E_j\) and the pixels on the prediction boundary \(\hat{E}_j\). For numerical fidelity and total
boundary length adjustment, set the weighted constant $\lambda_j$ and $\gamma_j$ is 1.

### 3.1 Cross-entropy Loss
We use the sum of two linear maps $(W_L, b_L)$ and $(W_G, b_G)$ to calculate local and global features. Use the softmax function to calculate the probability of each pixel that is prominent or not. The calculation formula is:

$$P = \frac{e^{w_L^TX_{i(i)}}+b_L^x+b_L^0}}{\sum_{s \in \{0,1\}} e^{w_L^TX_{i(i)}}+b_L^x+b_L^0}$$

The calculation formula of cross entropy loss is as follows:

$$G_j = \frac{1}{N} \sum_{i=1}^{N} \sum_{s \in \{0,1\}} (p(v_i) = s)(\log(\hat{p}(v_i = s)))$$

### 3.2 IoU Loss
The full name of IoU is the intersection and union ratio, which calculates the ratio of the intersection and union of the predicted boundary and the true boundary. And IoU boundary loss also has important applications in image segmentation and object detection [15], we use IoU boundary loss to approximate the boundary length to calculate the boundary loss. In order to calculate the boundary loss, we use the Sobel operator and the tanh activation function to approximate the magnitude of the saliency mapping gradient, the Sobel operator detects the edge magnitude, and the tanh activation function maps the saliency gradient to a probability range of $[0,1]$. The IoU boundary loss calculates the error between the real boundary and the predicted boundary. The boundary pixels use the Sobel operator and the tanh activation function. The activation function predicts the probability range of the saliency gradient $[0,1]$. The boundary loss calculation formula is as follows:

$$IoU_{Loss} = 1 - \frac{|E_j \cap \hat{E}_j|}{|E_j| + |\hat{E}_j|}$$

### 4. Experiment and Analysis

#### 4.1 Experimental details and data sets
In pre-training, we use VGG-16 as our backbone network for initialization, the learning rate is set to, the number of iterations is set to 10, and the training data set contains 2187 images.

We evaluated the performance of our method (SELNet) on four different public benchmark data sets, and all achieved good results. NJUD2000 contains 1985 pictures. We use 1487 NJUD2000 images for training, so the final test set of NJUD2000 contains 498 pictures. NLPR1000 contains 1000 images, we use 700 images for training, and the remaining 300 images for testing. The LFSD data set contains 100 test images. The SIP data set contains 80 test images.

#### 4.2 Evaluation Criteria
We use two evaluation indicators to evaluate the performance of our model and other salient object detection methods, including F-measure($F_\beta$) and Mean Absolute Error (MAE). The $F_\beta$, containing accuracy and recall rate corresponding to the PR curve, we calculate the corresponding binary mapping, and then calculate the accuracy/recall of all boundary mappings in the data set. The PR curve of the dataset explains the average accuracy and recall rate of the saliency images under different thresholds. F-measure is the harmonic average of average precision and average recall, which can be expressed as:
Where $\beta^2 = 0.3$, the larger the value, the better the quality of the predicted saliency map generated by its deep network. Mean Absolute Error (MAE) can be expressed as:

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

(10)

Where MAE is the average absolute difference between the predicted saliency map S and the true value map G in the pixel direction. Where W and H are the width and height of the specified predicted significance map S and truth map G.

4.3 Experimental results and Comparison

In the experimental model comparison, we selected nine models ACSD, SE, DF, CTMF, MMCI, TAN, D3Net, DANet, CoNet for comparison. We randomly select five pictures and these nine models for comparison experiments. The comparison of experimental prediction maps is shown in Figure 4.

As can be seen from Figure 4, the algorithm proposed in this paper can accurately locate the saliency object in the original image. Compared with the other nine models, the saliency map is closer to the true value map. Therefore, adding edge information to the model is beneficial to us. Accurately find the salient object in the original image. At the same time, in order to more intuitively see the algorithm comparison with the nine models, we use the two indicators of MaxF and MAE to compare with the other nine models. The results are shown in Table 1.

In order to better reflect the comparison between our model and other methods, we use the PR curve to illustrate. The PR curve includes precision precision and recall rate. Precision can be understood as the proportion of correct predictions among all positive examples of prediction (including positive examples of correct predictions and positive examples of incorrect predictions). The recall rate can be understood as Among all the real positive cases, the proportion that is correctly predicted. The wider the range of the PR curve, the better the performance. The PR curves of the four data sets are shown in Figure 5 (a), (b), (c), (d).
Table 1  Indicator comparison

| Indicator | NLPR1000 | NJUD2000 | LFSD   | SIP    |
|-----------|----------|----------|--------|--------|
|           | MaxF     | MAE      | MaxF   | MAE    | MaxF   | MAE    | MaxF   | MAE    |
| Ours      | 0.919    | 0.025    | 0.913  | 0.037  | 0.875  | 0.068  | 0.900  | 0.039  |
| ACSD      | 0.534    | 0.179    | 0.696  | 0.202  | 0.837  | 0.188  | 0.763  | 0.172  |
| SE        | 0.691    | 0.091    | 0.734  | 0.169  | 0.064  | 0.167  | 0.661  | 0.164  |
| DF        | 0.735    | 0.089    | 0.770  | 0.085  | 0.679  | 0.138  | 0.657  | 0.185  |
| CTMF       | 0.724    | 0.056    | 0.788  | 0.085  | 0.756  | 0.119  | 0.694  | 0.139  |
| MMCI       | 0.795    | 0.044    | 0.812  | 0.079  | 0.722  | 0.132  | 0.818  | 0.086  |
| TAN        | 0.796    | 0.041    | 0.844  | 0.061  | 0.771  | 0.111  | 0.830  | 0.075  |
| D3Net      | 0.897    | 0.030    | 0.900  | 0.046  | 0.834  | 0.058  | 0.861  | 0.063  |
| DANet      | 0.885    | 0.024    | 0.893  | 0.042  | 0.852  | 0.074  | 0.854  | 0.055  |
| CoNet      | 0.848    | 0.031    | 0.847  | 0.047  | 0.848  | 0.063  | 0.842  | 0.063  |

Fig.5  PR curve
5. Conclusion
In this article, we propose a fully convolutional neural network (SEANet) method based on dual-stream edge attention to guide local features. In the model, we add edge features for learning, improve the existing spatial attention model, and combine edge features input with the multi-scale features extracted from the VGG network to obtain edge perception features and introduce contrast. The edge perception features are enhanced, and finally the edge perception features and contrast features of the dual-stream output are integrated through upsampling and deconvolution to obtain the final prediction image. We compare with other nine models on four benchmark data sets, and the experimental results show that our method has achieved better detection results.

References
[1] Cheng M, Mitra N J, Huang X, et al. Global Contrast Based Salient Region Detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2015, 37(3):569-582.
[2] Yang C. Salient object detection based on prior fusion and manifold ranking[D]. Dalian University of Technology. 2016.
[3] Chen L C, Papandreou G, Kokkinos I, et al. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2018, 40(4):834-848.
[4] Li Q, He C, Chen Z X, Liu C Y. Bottom-up image salient object detection research[J]. Optoelectronics. Laser, 2016.
[5] Evan Shelhamer, Jonathan Long, Trevor Darrell. Fully Convolutional Networks for Semantic Segmentation[M]. IEEE Computer Society, 2017.
[6] He C, Chen Z X, Liu C Y. Bottom-up image salient object detection research[J]. Optoelectronics. Laser, 2016.
[7] Zhang J, Dai Y, Porikli F, et al. Deep Edge-Aware Saliency Detection[J]. 2017.
[8] Woo S, Park J, Lee J Y, et al. CBAM: Convolutional Block Attention Module[J]. 2018.
[9] Liu B X, Shi J M, Zhao D P, et al. Contrast analysis of object and background and its application[J]. Aerospace Electronic Warfare, 2008(03):48-51.
[10] Mumford D, Shah J. Optimal approximations by piecewise smooth functions and associated variational problems[J]. Communications on Pure & Applied Mathematics, 1989, 42(5):577-685.
[11] Brox T, Cremers D. On local region models and a statistical interpretation of the piecewise smooth Mumford-Shah functional[J]. International Journal of Computer Vision, 2009, 84(2):184-193.
[12] Zhu S C, Yuille A. Region competition: unifying snakes, region growing, and Bayes/MDL for multiband image segmentation[J]. IEEE Transactions on Pattern Analysis & Machine Intelligence, 1996, 18(9):416.
[13] Luo Z, Mishra A, Achkar A, et al. Non-local Deep Features for Salient Object Detection[C]/2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, 2017.
[14] Kang K, Wang X. Fully Convolutional Neural Networks for Crowd Segmentation[J]. Computer Vision and Pattern Recognition, 2014, 49(1):25-30.
[15] Popovic A, Matias de la Fuente, Engelhardt M, et al. Statistical validation metric for accuracy assessment in medical image segmentation[J]. International Journal of Computer Assisted Radiology & Surgery, 2007, 2(3-4):169-181.