ISSN 1746-7659, England, UK
Journal of Information and Computing Science
Vol. 17, No. 1, 2022, *pp*.028-039

CENet: Content-aware and Edge-aware Network for Salient Object Detection

Zhikuan Gao

School of Mathematics and Statistics, Nanjing University of Information Science & Technology,
Nanjing, 210044, China

(Received October 15, 2021, accepted January 13, 2022)

Abstract: Benefitting from Fully Convolutional Networks (FCNs), salient object detection methods have achieved prominent performance. However, there are still some challenges in this task: 1) lack of effective feature representation and integration make the result salient maps lose some regions of object, or bring some non-saliency regions. 2) suffering from the continuous pooling or stride operations, the predicted maps will lose some important spatial detail information, especially the boundary of object. To address these two problems, we propose the Content-aware and Edge-aware network (CENet) which contains three sub-modules: 1) we design a content-aware feature extraction module which uses a transformer block and channel-wise attention mechanism to capture the distinct content features and suppresses the non-saliency regions. 2) an edge-aware feature extraction module is introduced to learn the boundary features and predict the intact edge of the salient object. 3) a feature fusion module is proposed to integrate features from the first two module in a learning way. We also design a hybrid loss function which has better performance than the widely-used binary cross entropy loss. Results show that, our method can detect the intact salient object without losing regions of object or bring some non-saliency regions, and can also obtain the precise boundary. Experimented on several datasets, our method can achieve the state-of-art performance.

Keywords: Saliency object detection, Content-aware and edge-aware feature, Axial-attention transformer, Learning-based feature fusion one.

1. Introduction

Salient object detection aims at highlighting the most visually distinctive object in an image. Different from fixation prediction, which predicts the fixation points or locations that attract human attention mostly at firstly glance [1], salient object detection focuses on predict integrated object or regions and is used as a preprocessing step of many computer vision applications, such as semantic segmentation, image retrieval, image editing, image retargeting, person re-identification, video summarization , video salient object detection [2].

In the past few decades, lots of methods have been proposed for salient object detection. Conventional methods, which are motivated by the research on the human visual attention mechanism, usually use heuristic priors and hand-crafted feature such as color, contrast and texture. Although these methods performance well on some simple scenario, low-level features are not robust enough to distinguish the salient object and background regions in complex cases since the hand-crafted features have limit capability to capture the high-level semantic and structural information. Recently, the merge of Fully Convolutional Neural Networks (FCNs) [3] helps salience object detection make great strides and the FCNs becomes the widely used structure in salient object detection since its prominent capability to capture robust high-level features and stronger semantic information. Different from previous Convolutional Neural Networks (CNNs), FCN is an end-to-end structure which predicts pixel-wise outputs from arbitrary-size inputs at both learning and inference stage.

Motivated by this, more and more FCNs based methods were proposed and performed well on salient object detection task, however, there are still several problems[4-5]. To address above problems, we propose a content-aware and edge-aware network, referred to as CENet, for salient object detection. Our mainly contributions are summarized as follows:

- We present a novel network CENet for salient object detection, aiming to detect and segmentation the intact salient object with precise boundary. To obtain the intact object, we design a content-aware feature extraction module which learns more robust sematic information and more discriminative feature representation. To obtain the distinct edge, we design an edge-aware feature extraction module which transforms the HED edge maps into edge features and refine the boundary of predicted object.
- We propose a residual learning-based content-edge feature fusion module which integrate content features and edge features gradually. A batch concatenate operation is used during the fusion process to combine the different features wisely.
- We present a novel hybrid loss, which fuses binary cross entropy, focal loss, SSIM loss and IOU loss and has better performance than the widely used binary cross entropy loss, aiming to leads network to pay more attention on content and edge information.

2. Related Work

2.1. Traditional Methods

Early methods usually detect salient object using heuristic cues or hand-craft features, such as center prior, boundary and background, color contrast and texture. Besides, many graph-based methods are proposed for saliency detection. In [6], Shan et al. provide seeds for manifold using background weight map. In [7], Yang et al. construct a close-loop graph in which each node is a super pixel and propose a two-stage scheme. This way, the salient object detection problem is transformed to manifold ranking. Although these traditional approaches make great strides in salient object detection and performance well on simple scenario, they still hardly give a satisfying result in most complex scenes due to lack of the capability to capture high-level features.

2.2. FCN-based Methods

Recently, Convolutional Neural Network (CNN) has represented powerful capability of extracting high-level features and achieved prominent results in image classification. Motivated by this, some methods classify each pixel or super pixel based on CNN. They predict each pixel whether belongs to saliency regions or not. However, the saliency maps produced by these patch-wise based methods are usually coarse since the fully connected (fc) layers destroy the spatial structure of object.

To achieve pixel-wise prediction and obtain more precise results, in [3], Long et al. propose a Fully Convolutional Network (FCN) which is originally used to semantic segmentation. Several approaches begin to utilize the FCNs to generate end-to-end saliency map[10].

2.3. Edge-aware Guidance Methods

Existing methods can locate salient object precisely, however, how to segment the object from background with distinct boundary is still a challenge. To solve this problem, several methods detect salient object with edge-aware guidance[11-13].

2.4. Content-aware Guidance Methods

In addition to edge information, content information is also important for salient object detection. It helps the network comprehend what the object is and obtain the entire object. To capture more content information some methods enlarge the receptive field by ASPP instead of pooling operations to extract multi-scale context information However, with the dilation rate increasing, this will cause gridding problem that the local information is lost after the ASPP operation. Inspired by SENet [14], several methods introduce the attention mechanism to network architecture.

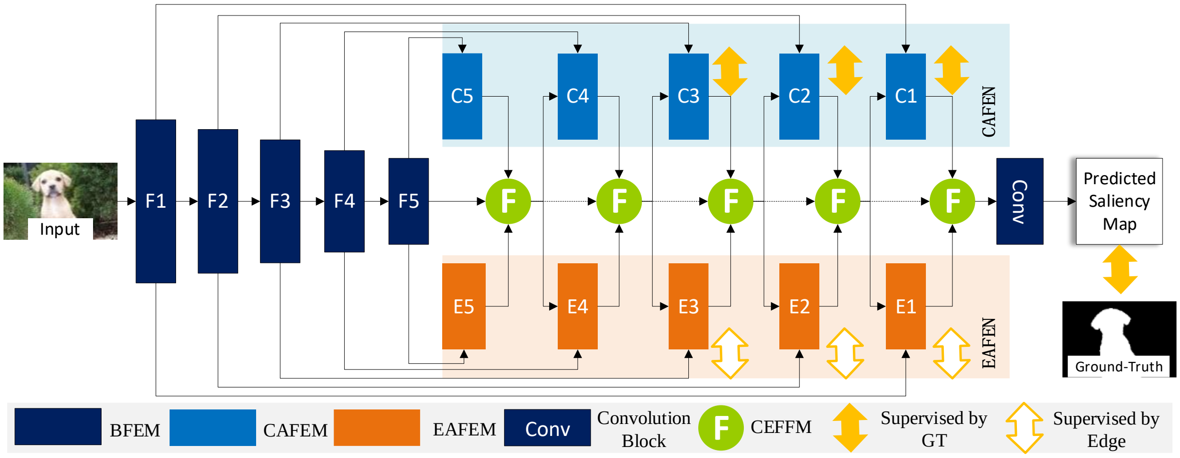

Figure. 1. Overview of the Content-aware and Edge-aware Network (CENet).

3. Proposed Method

To solve the problem of content inaccuracy and boundary blurry problem, we propose the Content-aware and Edge-aware Network, referred to as CENet, for salient object detection. The CENet contains a Backbone Feature Extraction Network (BFEN), a Content-Aware Feature Extraction Network (CAFEN), an Edge-Aware Feature Extraction Network (EAFEN), and a Content-Edge Feature Fusion Network (CEFFN). Besides, we also design a novel hybrid loss function and a multi-path optimizing strategy for supervising the learning of content and edge specially. When the network gets an input image, the BFEN extracts multi-scale and multi-level features firstly. Then the CAFEN and the EAFEN extract useful content and edge information from these features respectively and the content features and edge features at each level are fused by CEFFN. During the learning stage, we use the defined loss function to supervise the final saliency map and side-out maps which include content maps and edge maps. We use a multi-path optimizing strategy to update parameters selectively during training stage that can lead the network to pay more attention to content and edge information. Figure.1. shows the overview of the proposed method.

3.1. Backbone Feature Extraction Module

Similar to most salient object detection methods, we select VGG-16 as the backbone network to extract hierarchical multi-level features for the further step work. However, the VGG-net and Resnet are designed for image classification task originally, and thus we need to do some special modification for salient object detection task. The feature extraction module has total five stages. For VGG-16, we cut the last three fully connection layers since we need an end-to-end output and the fc layers lose too much spatial structure information. We use the last layer of each convolution block as the side-out features denoted as $F_i (i \in \{1,2,3,4,5\})$. To be specific, the feature set $F = \{F_i\}_{i=1}^{5}$ is the set of output of Conv1_2, Conv2_2, Conv3_3, Conv4_3 and Conv5_3 from VGG-16. For Resnet-34, inspired by [45], we set the kernel size of the input convolution layer which has 64 convolution filters to 3×3 and set the stride to 1 instead of size of 7×7 and stride of 2.

3.2. Content-Aware Feature Extraction Module

The Content-Aware Feature Extraction Network has five stages which are symmetrical to the five stages in BFEN. Each stage of CAFEN consists of a basic decode block and a CAFEM. The basic decode block consists of three convolution layers followed by a batch normalization layer and a ReLU activation function. The input of each stage is the concatenated feature maps of the upsampled output from its previous Content-Edge Feature Fusion Module (CEFFM) and its responding stage in the BFEN. Then the output feature maps of basic decode block are fed into CAFEM to extract the content features. Denote content feature extracted by each stage as $C_i\,(i \in \{1,2,3,4,5\})$, We show the process in Figure.2.

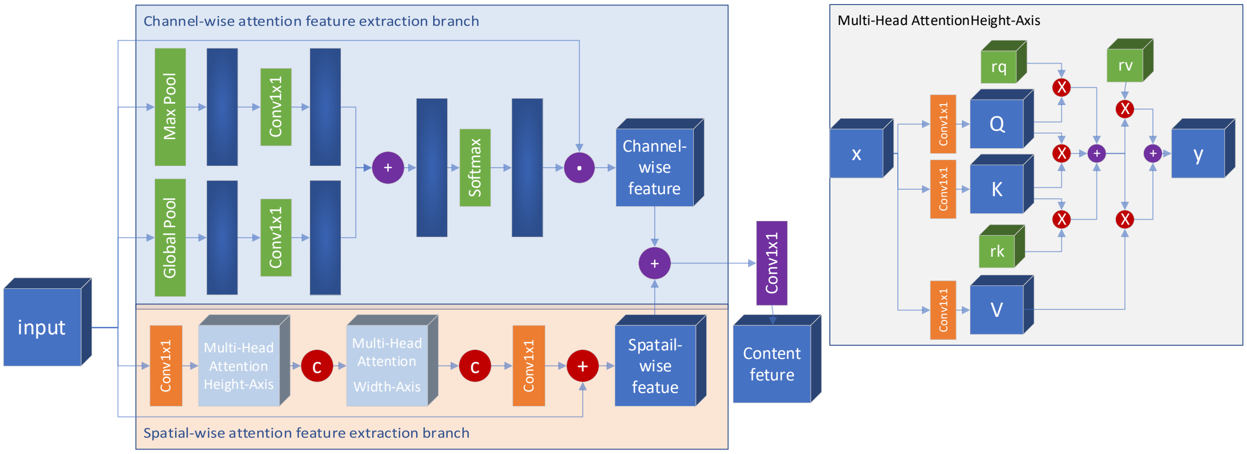
Channel-wise attention feature extraction branch
Max Pool
Conv1x1
Global Pool
Conv1x1
+
Softmax
•
Channel-wise feature
+
Conv1x1
input
Conv1x1
Multi-Head Attention Height-Axis
C
Multi-Head Attention Width-Axis
C
Conv1x1
+
Spatail-wise featue
Content feture
Spatial-wise attention feature extraction branch
Multi-Head AttentionHeight-Axis
x
Conv1x1
Conv1x1
Conv1x1
rq
Q
K
rk
V
rv
X
X
X
X
X
+
+
y

Figure.2. The illusion of content-aware feature extraction module (CAFEM).

3.2.1. Channel-wise Attention Feature Extraction Branch

To be specific, we generate channel-wise statistics by using both global average pooling operation and max pooling operation to model the channel dependence relationship and exploit different semantics information with a global receptive field. By global average pooling and max pooling, feature map with width W and height H in one channel is transformed to a real value and the whole feature maps $F_i \in \mathbb{R}^{W\times H\times C}$ with width W, height H and channels C is transformed to a vector $G \in \mathbb{R}^{1\text{x}1\times C}$ and a vector $M \in \mathbb{R}^{1\text{x}1\text{x}C}$. It can be described as:

$$G = \mathrm{GP}(x) = \frac{1}{W \times H} \sum\nolimits_{j=1}^{W} \sum\nolimits_{k=1}^{H} x$$

$$M = \mathrm{MP}(x) = \max(x)$$

where $[j,k]$ is the pixel location, $\mathrm{GP}(\bullet)$ d $\mathrm{MP}(\bullet)$ denote global average pooling and max pooling.

we use a 1D convolution layer without dimensionality reduction operation after global average pooling and max pooing to learn the non-linear and non-mutual relationship between channels. Besides, to appropriately model local channel-wise dependence for convolution layer with different channel numbers, we select the kernel size of convolution adaptively. The adaptive kernel size denoted as k is related to the channel number C and the relationship between k and C is:

$$k = \left| \frac{\log_2 C}{\gamma} + \frac{b}{\gamma} \right|_{odd}$$

Where γ and C are two hyper parameters, $|\bullet|_{odd}$ denotes the nearest odd number. We set γ and C to 2 and 1. Then the outputs are integrated by element-wise addition operation. After that, a SoftMax is used to normalize the value of vector to [0, 1] and we obtain the channel-wise attention weight vector which represent the importance weight of different channels. The large weight of a pixel means that this pixel has more useful information and is more likely the saliency pixel. On the contrary, if the weight is small, this pixel may be the background pixel. Finally, the original features maps dot with the weight vectors in channel dimensionality to obtain the channel-wise attention features. We refer to $F_i \in \mathbb{R}^{W\times H\times C}$ and $C_i^{\text{channel}} \in \mathbb{R}^{W\times H\times C}$ as the output of basic decode convolution block and the channel-wise attention features respectively. The channel-wise attention feature extraction branch can be described as:

$$C_i^{\text{channel}} = \sigma(\text{conv}(\text{GP}(F_i), w_g) \oplus \text{conv}(\text{MP}(F_i), w_m)) \odot F_i$$

where $\sigma(\bullet)$ represents SoftMax activation function. $\text{conv}(\bullet)$ the 1×1 convolution. w_g and w_m are the weights of two convolution layers, $\oplus$ and $\odot$ represent element-wise addition operation and channel-wise dot multiplication operation respectively.

3.2.2 Spatial-wise Attention Feature Extraction Branch

Global context information is important for segmenting the intact salient object, however, the local features extracted by convolution lack global dependence relationship in spatial domain since the convolution is a local operation. As a result, the learned features are not powerful enough to discriminate the salient object and background. For the sake of capturing global information and modeling the long-range dependence relationship between pixels, we introduce axial-attention transformer proposed by [20] into the CAFEM. The axial-attention is an improvement of self-attention mechanism. Firstly, it introduces positional embedding on the basis of self-attention to obtain more efficient positional relationship between different pixels. It conducts the computation and information extraction along height-axis and width-axial respectively and thus the complexity of computation is reduced from quadratic to linear.

Given an input feature map $X \in \mathbb{R}^{H\times W\times C}$, we first use three 1×1 conv to transform X with linear mapping and obtain three important elements in transformer: query l, key $K \in \mathbb{R}^{H\times W\times C/2}$ and value $V \in \mathbb{R}^{H\times W\times C}$. The linear mapping functions are $Q = W_Q X$, and $V = W_V X$, where $W_Q \in \mathbb{R}^{C\times C/2}$, $W_K \in \mathbb{R}^{C\times C/2}$ and $W_V \in \mathbb{R}^{C\times C/2}$ are the learnable matrices. To capture diverse information, we adopt multi-head attention mechanism and split Q, K V to 8 heads $\{Q_i\}_{i=1}^{8}$, $\{K_i\}_{i=1}^{8}$ and $\{V_i\}_{i=1}^{8}$ respectively. An output $y_o \in \mathbb{R}^{C\times C/2}$ at position $o = (i, j)$ in a head of width-axial attention is computed as:

$$y_o = \sum\nolimits_{p\in N_{1\times m(o)}} \text{Soft}\max\nolimits_p (q_o^T k_p + q_o^T r_{p-o}^q + k_p^T r_{p-o}^k)$$

where N the whole location lattice $1\times m(o)$ is local $1\times m$ region centered around location $o = (i, j)$. r^q, r^k and r^v denote the learnable positional embedding matrix for Q, K and V respectively. $\text{Soft}\max_p$ is SoftMax function applied to all possible $p = (a, b)$ positions.

For the whole spatial-wise attention feature extraction branch, we use a ResNet block embedded with height-axial attention and width-axial attention. Given the input feature map $F_i \in \mathbb{R}^{H\times W\times C}$ with height H width W and channels C, we first use a 1×1 convolution reduce the channels to $H \times W \times C/2$ and compute multi-head attention along height-axial. We concatenated these output features in channel-wise and then compute multi-head attention along width-axial. The output features are also concatenated after attention computation. Finally, we use 1×1 convolution to restore the channels to $H\times W\times C$ and add the feature map with the original feature F_i element-wise addition. In this paper, we set the head number of multi-head attention to 8. The m in width-axis attention is the width of input and in height-axis is the height of input. The spatial-wise attention feature extraction branch is formulated as:

$$C_i^{\text{spatial}} = \text{Conv}(\text{Cat}(\{A_{\text{width}}^k(\{A_{\text{width}}^j(\text{conv}(F_i))\}_{j=1}^{8})\}_{k=1}^{8})) \oplus F_i$$

where A_{width}^k, A_{width}^j denote the k th head of width-axis attention and the j th head of height-axis attention

By these two branches, we obtain channel-wise attention feature C_i^{channel} and spatial-wise attention feature C_i^{spatial} which extract more powerful global context information and semantic information from two aspects. Finally, we use an element-wise addition and a 1×1 convolution to integrate the features from these two branches. The content feature is:

$$C_i = \text{conv}(C_i^{\text{spatial}} \oplus C_i^{\text{channel}})$$

3.3. Edge-Aware Feature Extraction Module

Edge feature is another important feature for salient object detection task. However, due to consecutive pooling operation and up-sampling operation with large factor, lots of detail information is lost and the boundary of predicted salient object becomes incomplete and unsharp. Thus, supplementing with detail information, especially efficient boundary information, can greatly improve the accuracy of the results and obtain more complete salient object with distinct boundary. To extract more useful edge features, inspired by [58], we design an Edge-Aware Feature Extraction Module (EAFEM). This module has two branches: one is edge prior extraction branch which extracts the prior knowledge of edge as the heuristic cues for further steps; another is edge transform branch which transforms the intermediate features extracted from BFEN to edge feature with the edge prior extracted by edge prior extraction branch. Figure.3. shows the details of EAFEM.

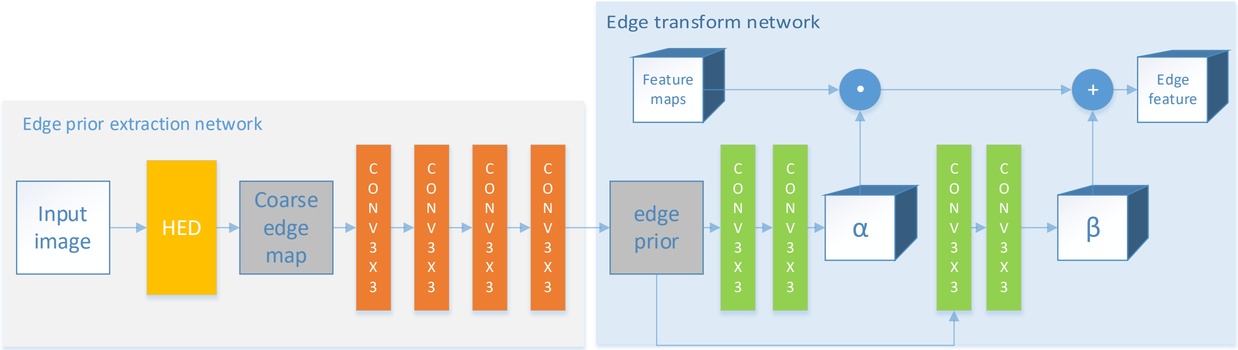

Figure.3. The illusion of edge-aware feature extraction module (EAFEM).

In edge prior extraction branch, we use holistically nested edge detection (HED) Network to detect the edge of input image and obtain the coarse edge map e_{coarse}:

$$e_{\text{coarse}} = \text{HED}(I)$$

where $\text{HED}(\bullet)$ is the holistically nested edge detection network described in [19] and I the input image. Then we add four convolutional layers with kernel size of 3×3 and strides of 1 to get the edge prior condition $e_{\text{prior}} \in \Psi$. Each convolutional layer is followed by a ReLU activation function. In the edge transform branch, we learn a mapping function which outputs a modulatory parameter pair (α, β):

$$\varphi : \Psi \mapsto (\alpha, \beta), \qquad (\alpha, \beta) = \varphi(\Psi)$$

Then we transform the intermedia feature maps to edge features E_i with the modulatory parameter pair:

$$E_i = F_i \odot \alpha \oplus \beta$$

$$\alpha = \text{ReLU}(\text{conv}_{3*3}(\text{ReLU}(\text{conv}_{3*3}(e_{\text{prior}}))))$$

$$\beta = \text{ReLU}(\text{conv}_{3*3}(\text{ReLU}(\text{conv}_{3*3}(e_{\text{prior}}))))$$

where F_i the intermedia feature maps extracted from BFEN. Similar to CAFEN, we extract the hierarchical edge features $E_i, i = 1,2,3,4,5$ from five backbone feature $F_i, i = 1,2,3,4,5$ by the EAFEN. Besides, the prior features extracted by edge prior extraction network are shared in EAFEM. In other words, we only extract edge prior from input image once and use this prior during the edge transformation of different level features.

3.4. Content-Edge Feature Fusion Module

After obtaining different level content features and edge feature, we need to find an effective way to aggerate these two kinds of learned features. We design a content-edge feature fusion module (CEFFM) which enhances the features in a residual learning way. As shown in Figure. 4., we fuse the C_i and E_i by pixel-wise addition to enhance both the content-aware features and edge-aware features. Then we add three convolutional layers with kernel size of 3×3 and strides of 1 and each of which is followed by a batch normalization layer and a ReLU activation function. Finally, we obtain the fused feature maps S_i which contain rich context and edge information by adding the up-sampled features S_{i+1} from level $i+1$. The total fusion process is formulated as follow:

$$S_i = \mathrm{UP}(S_{i+1}) \oplus T_{\mathrm{conv}}(C_i \oplus E_i)$$

where the $T_{\mathrm{conv}}(\bullet)$ present the consecutive convolution, BN and ReLU operations.

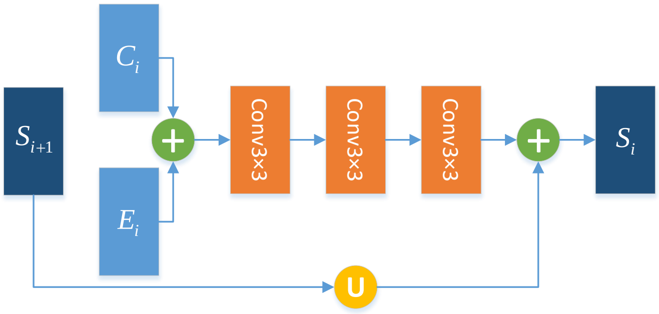

Figure.4. The illusion of feature fusion module.

3.5. Hybrid Loss

In this paper, we design a novel hybrid loss function for training the network which leads to better result on content and boundary of salient object. We design this loss function in three perspectives: content map, edge map and fusion map corresponding to content loss L_C, edge loss L_E and fusion loss L_F. These loss functions are formulated as:

$$L_C(P,T) = L_{iou}(P,T) + L_{ssim}(P,T)$$

$$L_E(P,T) = L_{cbce}(P,T_E)$$

$$L_F(P,T) = \alpha_1(L_{iou}(P,T) + L_{ssim}(P,T)) + \alpha_2 L_{cbce}(P,T)$$

where L_{cbce} presents the Class-Balanced Cross-Entropy Loss. L_{iou} the Intersection Over Union Loss. L_{ssim} notes Structural Similarity Loss. α_1 and α_2 note the weights of two portion in fusion loss and we set them to 0.6 and 0.4 by experiment.

IOU Loss is originally the evaluation measure of object category segmentation problem and object detection task. The IOU compares the similarity of ground-truth mask and prediction map to evaluate the quality of results. The IOU loss is defined as:

$$L_{iou}(P,T) = 1 - \frac{2|P \cap T|}{|P| + |T|} = 1 - \frac{\sum_{i=1}^{W}\sum_{j=1}^{H} P(i,j)T(i,j)}{\sum_{i=1}^{W}\sum_{j=1}^{H}[P(i,j) + T(i,j) - P(i,j)T(i,j)]}$$

where (i,j) the location of pixel, P presents the predicted salience map generated and T is the ground-truth saliency mask. We hope the confidence of object is higher and the confidence of background is lower. By optimizing the IOU loss, the pixel of background will be penalized and the loss of predicted salient object tends to decrease to zero. Thus, the IOU loss will encourage the CAFEN to focus more on context information.

SSIM loss considers the structural information of neighborhood patch at pixel level. The SSIM loss is defined as:

$$L_{ssim} = 1 - \frac{(2\mu_x\mu_y + c_1)(2\sigma_{xy} + c_2)}{(\mu_x^2 + \mu_y^2 + c_1)(\sigma_x^2 + \sigma_y^2 + c_2)},$$

where μ_x, μ_y represent the mean intensity of train image and binary ground-truth mask respectively, σ_x and σ_y are the standard deviation which is the unbiased estimate in discrete form, σ_{xy} is covariance and we set $c_1 = 0.01^2$ and $c_2 = 0.03^2$ to avoid dividing by zero. In this paper, we use IOU loss and SSIM loss to optimize the network during learning the content-aware features.

For one image, the edge pixel numbers and non-edge pixel numbers of ground-truth map are extremely imbalanced. In this case, using the binary cross entropy loss may lead to imprecise result. Thus, we introduce class-balanced cross entropy for optimizing the EAFEN. We use a class-balancing weight β eliminate imbalance between edge and non-edge. The class-balanced cross-entropy loss is defined as:

$$L_E(P,T_E) = -\beta\sum_{i\in E_+} T_{E_i}\log(P_i) - (1-\beta)\sum_{i\in E_-}(1-T_{E_i})\log(1-P_i),$$

where $\beta = |T_{E-}|/|T_E|$, $1-\beta = |T_{E+}|/|T_E|$. $|T_{E-}|$ and $|T_{E+}|$ are the edge label set and non-edge label set of ground-truth set respectively.

4. Experimental Results

4.1. Implement details

We utilize DUTs-TR dataset which contains 10553 images to train our proposed network. During the training stage, to avoid overfitting, we adopt several data augment approaches includes random rotation, horizontal flip and random cropping. The input images are first resized to 256×256 and then we cropping them to 224×224 randomly. The parameters in BFEN are initialized by pretrained VGG-16 or ResNet-34. The other parameters are initialized from normal distribution randomly. Our network is training by AdamW optimizer with beta=(0.9, 0.999), eps=1e-6, weight decay=0. We set the initial learning rate to 1e-5 which is divided by 10 if no improvement in training loss.

4.2. Datasets and evaluation metrics

4.2.1 Datasets

We evaluate our proposed methods on six challenging benchmark datasets: ECSSD, SOD, DUTs,. These datasets of which images have complex background, multiply salient objects and multi-scale object can evaluate performance of our proposed method effectively. ECSSD has 1,000 structurally intricate images with rich semantic information and complex background contents. DUTs is a large dataset which consist of 10553 training images selected from ImageNet DET train and validation set and 5019 test images selected from ImageNet test dataset and SUN dataset.

4.2.2 Evaluation metrics

We use six metrics to evaluate the performance of our method as well as previous state-of-the-art saliency detection methods, namely Precision-Recall (PR) curves, F-measures curves, weighted F-measure score, E-measure score, Mean Absolute Error (denoted as MAE) score, and Structural measure (S-measure).

The PR-curve depicts the relationship between precision rate and recall rate in binary classification and it is calculated as:

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TN}+\text{FN}}$$

where T and F note the predicted pixel is whether saliency pixel. F and N note the pixel belongs to salient object or background in ground-truth mask. TP presents true-positive which means the predicted salient pixel is also that in ground-truth mask. TN denotes true-negative which computes how many predict salient pixels are the background pixels in ground-truth mask. FP and FN denote false-positive and false-negative respectively. Setting different threshold may get different binary mask. We select integers from 0 to 255 as threshold, each of which is applied to get a binary mask. The PR-curve is plotted by calculating the Precision / Recall value on each threshold.

F-measure score is calculated by considering both the aforementioned precision and recall rate:

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

where β the weight which decides to give more focus on precision or recall. It is usually set to 0.3 to increase the importance of precision. We also calculate the maximal F-measure score (max F) and the average F-measure score (avg F) to evaluate the performance of our proposed model.

Weighted F-measure aims to explore the dependency between pixel and its neighborhood by modifying the TP, TN, FP, FN, to non-binary values. Besides, different errors will be attributed different importance. The Weighted F-measure is defined as:

$$F_{\beta}^{\omega} = \frac{(1+\beta^2)\text{precision}^{\omega}\times\text{recall}^{\omega}}{\beta^2\text{precision}^{\omega}+\text{recall}^{\omega}},$$

the detail calculation procedure of weighted precision and weight recall is depicted in [21]

Mean Absolute Error is the average of the absolute error between ground-truth mask and predicted saliency map:

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

where W and H note the width and height of image, $G(i,j)$ and $S(i,j)$ denote the value of pixel at location (i,j) in ground-truth and predicted saliency map respectively.

Structural measure considers both the object-aware and region-aware structural similarity between ground-truth mask and predicted saliency map:

$$S = \alpha * S_o + (1-\alpha) * S_r$$

,

where S_o and S_r represent object-aware and region-aware structural similarity respectively. α is usually set to 0.5.

Enhanced-alignment measure considers both the pixel-wise errors and image-wise errors by integrating the local pixel and global statistical information. It can be calculated as:

$$Q_S = \frac{1}{W\times H}\sum_{i=1}^{W}\sum_{j=1}^{H}\phi_S(i,j)$$

where ϕ_S is the enhanced alignment matrix described in [22].

4.3. Comparison with the state-of-the-art

We compare our proposed CENet with the other 11 state-of-art salient object detection methods including DHS [8], ELD [23], DSS [24], Amulet [10], DLS [25], PiCANet [26], C2SNet [27], PAGRN [28], AFNet [11], BASNet [12], F3Net [16].To ensure the fairness of comparison, we obtain saliency maps from public webpages of the authors or by running their fine-tuned model.

Table 1 shows the detail evaluation result. To be specific, first, we select VGG-16 as backbone and we use two convolution layers instead of the last two fully connection layers for end-to-end output. As we can see, simple VGG-16 cannot detect precise result since lack of the capability of extracting efficient features. Then, we add CAFEN to the backbone and the performance is improved since more powerful content information are extracted by CAFEN. Next, we add EAFEN on the basis of backbone+CAFEN and we use element-wise addition to integrate the feature maps from CAFEN and EAFEN. The extracted edge feature maps contribute to the precise result although the performance is improved slightly. Finally, to prove the effectiveness of learning-based feature fusion, we add CEFFN on the basis of backbone+CAFEN+EAFEN and achieve the best performance with the help of CEFFN.

5. Conclusion

In this paper, we propose a novel content-aware and edge-aware network, denoted as CENet, for salient object detection. To capture more distinctive and more powerful features, we design several sub-modules. For content problem, we design a CAFEN to extract content information from multi-scale and multi-level backbone features. For edge problem, we design an EAFEN to transform the backbone features to edge features. Considering the complementary relationship between content feature and edge feature, we design a residual learning-based feature fusion module CEFFN to integrate features from CEFEN and EFEN. Moreover, we proposed a novel hybrid loss which combines class-balanced cross-entropy, IoU loss and SSIM loss to guide the network to get more precise results. Experimental results on six datasets show that the proposed method achieves competitive results under different evaluation metrics. In the future, we will consider to reduce the computational cost and compress model, and explore weakly-supervised learning method.

Table 1. Performance comparison between the proposed method and eleven state-of-the-art methods on six benchmark datasets in terms of the max F measure, MAE and S measure.

Method	Year	ECSSD						DUTS-TE						SOD					
		Fmax	avgF	MAE	wF	Sm	Em	Fmax	avgF	MAE	wF	Sm	Em	Fmax	avgF	MAE	wF	Sm	Em
MCDL	2015	0.836	0.795	0.100	0.727	0.802	0.864	0.669	0.590	0.105	0.535	0.712	0.776	0.433	0.320	0.342	0.239	0.412	0.540
LEGS	2015	0.827	0.785	0.118	0.688	0.786	0.845	0.652	0.580	0.137	0.509	0.696	0.757	0.397	0.301	0.355	0.238	0.421	0.535
MDF	2015	0.831	0.806	0.105	0.705	0.776	0.846	0.708	0.602	0.113	0.507	0.727	0.768	0.417	0..304	0..342	0.233	0.412	0.509
ELD	2016	0.868	0.816	0.078	0.785	0.841	0.883	0.737	0.625	0.092	0.608	0.753	0.785	0.422	0.346	0.345	0.281	0.433	0.549
DHS	2016	0.906	0.871	0.058	0.840	0.883	0.911	0.807	0.720	0.067	0.697	0.817	0.841	0.424	0.332	0.345	0.265	0.418	0.537
DCL	2016	0.900	0.875	0.067	0.820	0.868	0.903	0.781	0.678	0.087	0.606	0.795	0.816	0.429	0.337	0.348	0.263	0.420	0.542
DSS	2017	0.920	0.904	0.051	0.872	0.882	0.912	0.825	0.788	0.056	0.754	0.823	0.881	0.406	0.332	0.339	0.273	0.434	0.528
DLS	2017	0.852	0.821	0.085	0.772	0.806	0.865	-	-	-	-	-	-	-	-	-	-	-	-
Amulet	2017	0.915	0.868	0.058	0.840	0.894	0.901	0.777	0.677	0.084	0.658	0.803	0.793	0.414	0.346	0.356	0.286	0.425	0.539
SRM	2017	0.917	0.892	0.054	0.852	0.894	0.916	0.826	0.752	0.058	0.721	0.835	0.860	0.434	0.335	0.341	0.255	0.414	0.527
RADF	2018	0.923	0.904	0.048	0.882	0.893	0.923	0.819	0.770	0.061	0.747	0.825	0.862	0.402	0.333	0.345	0.277	0.432	0.531
PAGRN	2018	0.926	0.894	0.060	0.833	0.889	0.914	0.853	0.783	0.055	0.724	0.838	0.880	0.416	0.304	0.337	0.232	0.415	0.517
PiCANet	2018	0.931	0.884	0.046	0.865	0.913	0.910	0.851	0.749	0.054	0.746	0.860	0.851	0.444	0.345	0.343	0.283	0.427	0.524
C2S	2018	0.911	0.865	0.053	0.854	0.895	0.914	0.810	0.717	0.061	0.717	0.831	0.846	0.422	0.333	0.350	0.269	0.425	0.535
AFNet	2019	0.935	0.907	0.041	0.886	0.913	0.918	0.862	0.792	0.045	0.784	0.866	0.878	-	-	-	-	-	-
BASNet	2019	0.942	0.879	0.037	0.903	0.916	0.920	0.859	0.791	0.047	0.802	0.865	0.884	0.418	0.318	0.339	0.285	0.442	0.522
SSNet	2019	0.915	0.883	0.044	0.854	0.863	0.869	0.823	0.764	0.047	0.706	0.767	0.793	0.370	0.273	0.316	0.232	0.389	0.441
CENet	**2020**	**0.945**	**0.924**	**0.033**	**0.912**	**0.924**	**0.927**	**0.891**	**0.839**	**0.035**	**0.834**	**0.888**	**0.901**	**0.418**	**0.318**	**0.339**	**0.285**	**0.442**	**0.522**

6. Reference

[1] W. G. Wang, J. B. Shen, X. P. Dong, and A. Borji, *Salient object detection driven by fixation prediction*. Proc.

of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1711-1720. 2018.

[2] D. P. Fan, W. G. Wang, M. M. Cheng, J. B. Shen. *Shifting more attention to video salient object detection*, Proc. of the IEEE conference on computer vision and pattern recognition, pp. 8554-8564. 2019.

[3] J. Long, E. Shelhamer, and T. Darrell. *Fully convolutional networks for semantic segmentation*, Proc. of the IEEE conference on computer vision and pattern recognition, pp. 3431-3440. 2015.

[4] L. C. Chen, G. Papandreou, I. Kokkinos, K. Murphy, and A. Yuille. *Deeplab: Semantic image segmentation with deep convolutional nets, atrous convolution, and fully connected crfs*, IEEE transactions on pattern analysis and machine intelligence 40, no. 4 (2017), pp. 834-848.

[5] K. Simonyan, Karen, and A. Zisserman. *Very deep convolutional networks for large-scale image recognition*, arXiv preprint arXiv: (2014), 1409.1556.

[6] D. J. Shan, X. W. Zhang, and C. Zhang. *Visual saliency based on extended manifold ranking and third-order optimization refinement*, Pattern Recognition Letters 116 (2018): 1-7.

[7] C. Yang, L. H. Zhang, H. C. Lu, X. Ruan, and M. H. Yang. *Saliency detection via graph-based manifold ranking*, Proc. of the IEEE conference on computer vision and pattern recognition, pp. 3166-3173. 2013.

[8] N. Liu, and J. Han. *Dhsnet: Deep hierarchical saliency network for salient object detection*, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 678-686. 2016.

[9] Q. B. Hou, M. M. Cheng, X. W. Hu, A. Borji, Z. W. Tu, and P. Torr. *Deeply supervised salient object detection with short connections*, In Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3203-3212. 2017.

[10] P. P. Zhang, D. Wang, H. C. Lu, H. Y. Wang, and X. Ruan. *Amulet: Aggregating multi-level convolutional features for salient object detection*, Proc. of the IEEE International Conference on Computer Vision, pp. 202-211. 2017.

[11] M. Y. Feng, H. L. Lu, and E. Ding, *Attentive feedback network for boundary-aware salient object detection*, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1623-1632. 2019.

[12] X. B. Qin, Z. C. Zhang, C. Y. Huang, C. Gao, M. Dehghan, and M. Jagersand, *Basnet: Boundary-aware salient object detection*, Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7479-7489. 2019.

[13] J. X. Zhao, J. J. Liu, D. P. Fan, Y. Cao, J. F. Yang, and M. M. Cheng, *EGNet: Edge guidance network for salient object detection*, Proc of the IEEE International Conference on Computer Vision, pp. 8779-8788. 2019.

[14] J. Hu, L. Shen, and G. Sun. *Squeeze-and-excitation networks*, Proc. of the IEEE conference on computer vision and pattern recognition, pp. 7132-7141. 2018.

[15] S. Mohammadi, M. Noori, A. Bahri, S. G. Majelan, and M. Havaei, *CAGNet: Content-Aware Guidance for Salient Object Detection*, Pattern Recognition (2020): 107303.

[16] J. Wei, S. H. Wang, and Q. M. Huang, *F3Net: Fusion, feedback and focus for salient object detection*, arXiv preprint arXiv, (2019):1911.11445.

[17] X. Xu, J. X. Chen, H. D. Zhang, and G. Q. Han, *Dual pyramid network for salient object detection*, Neurocomputing 375 (2020): 113-123.

[18] S. H. Chen, X. L. Tan, B. Wang, H. C. Lu, X. L. Hu, and Y. Fu, *Reverse Attention-Based Residual Network for Salient Object Detection*, IEEE Transactions on Image Processing 29 (2020): 3763-3776.

[19] S. N. Xie, and Z. W. Tu, *Holistically-nested edge detection*, Proc. of the IEEE international conference on computer vision, pp. 1395-1403. 2015.

[20] H. Y. Wang, Y. K. Zhu, B. Green, H. Adam, A. Yuille, and L. Chen, *Axial-deeplab: Stand-alone axial-attention for panoptic segmentation*, Proc. of European Conference on Computer Vision, pp. 108-126. Springer, Cham, 2020.

[21] R. Margolin, L. Zelnik-Manor, and A. Tal. *How to evaluate foreground maps?*, Proc. of the IEEE conference on computer vision and pattern recognition, pp. 248-255. 2014.

[22] D. P. Fan, C. Gong, Y. Cao, B. Ren, M. M. Cheng, and A. Borji. *Enhanced-alignment measure for binary foreground map evaluation*, arXiv preprint arXiv, (2018), 1805.10421.

[23] Li G, Yu Y, *Deep contrast learning for salient object detection*, Proc of the IEEE Conf. Comput. Vis. Pattern Recognit 478–487.

[24] Hou Q, Cheng M M, Hu X, et al, Deeply supervised salient object detection with short connections[C]//Proc of the IEEE Conf. Comput. Vis. Pattern Recognit. 2017: 3203– 3212.

[25] Hu P, Shuai B, Liu J, et al. Deep level sets for salient object detection[C]Proc of the IEEE Conf. Comput. Vis. Pattern Recognit. 2017: 540–549.

[26] Liu N, Han J, Yang M H. Picanet: Learning pixel-wise contextual attention for saliency detection[C]//Proc of the IEEE Conf. Comput. Vis. Pattern Recognit. 2018: 3089–3098.

[27] Li X, Yang F, Cheng H, Liu W, et al. Contour knowledge transfer for salient object detection[C]//Proc of the European Conference on Computer Vision (ECCV). 2018: 355-370.

[28] Zhang X, Wang T, Qi J, et al. Progressive attention guided recurrent network for salient object detection[C]//Proc of the IEEE Conf. Comput. Vis. Pattern Recognit. 2018: 714– 722.