Local Context Attention for Salient Object Segmentation

Jing Tan, Pengfei Xiong, Zhengyi Lv, Kuntao Xiao, and Yuwen He

Megvii Technology Limited
{tanjing, xiongpengfei, lvzhengyi, xiaokuntao, heyuwen}@megvii.com

Abstract. Salient object segmentation aims at distinguishing various salient objects from backgrounds. Despite the lack of semantic consistency, salient objects often have obvious texture and location characteristics in local area. Based on this priori, we propose a novel Local Context Attention Network (LCANet) to generate locally reinforcement feature maps in a uniform representational architecture. The proposed network introduces an Attentional Correlation Filter (ACF) module to generate explicit local attention by calculating the correlation feature map between coarse prediction and global context. Then it is expanded to a Local Context Block(LCB). Furthermore, a one-stage coarse-to-fine structure is implemented based on LCB to adaptively enhance the local context description ability. Comprehensive experiments are conducted on several salient object segmentation datasets, demonstrating the superior performance of the proposed LCANet against the state-of-the-art methods, especially with **0.883 max F-score** and **0.034 MAE** on DUTS-TE dataset.

1 Introduction

Salient object segmentation aims at locating the most obvious and salient objects from a given image. It has been widely used in various challenging fields like automatic focus, autonomous driving, scene understanding, image editing, etc. In the past decades, salient object segmentation approaches [1–12] have already obtained promising performances on various benchmarks [13–17]. Nevertheless, most of the previous salient object segmentation methods treat it as a general semantic segmentation problem, which improves the performance by increasing the semantic receptive field, optimizing the edge accuracy or other methods.

As a pixel-level classification problem, the intrinsic properties of salient object segmentation determine that it is different from traditional semantic segmentation. Salient objects usually do not belong to the same category, and their textures and shapes are various, which makes it hard to distinguish the salient objects and the background by simply increasing the receptive field.

In contrast, salient objects often have obvious local context characteristic of the image. Despite their various sizes and locations, their vivid texture are always quite different from the surrounding backgrounds. Based on this prior, we believe that the calculation of each pixel attention on an equal basis is computationally
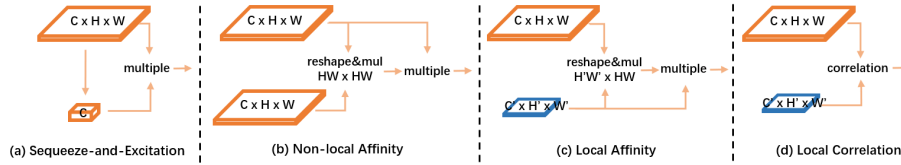
Fig. 1. Comparison of different attention approaches. From left to right: (a) Squeeze-and-Excitation[18]. (b) Non-local Affinity[19]. (c) Local Affinity and (d) Local Correlation Attention. It can be noticed that, Local Correlation is operated on the local context of global feature, instead of the entire feature map.

inefficient and out of focus in the previous approaches. Therefore, we rethink salient object segmentation task from a more macroscopic point of view, which separates the salient object from the distant irrelevant background, and then extracts local context features related to the object as supervision to enhance the distinguish ability of foreground object.

An intuitive way to extract local context features is to construct a coarse-to-fine architecture with multi-scale inputs[16, 20]. It generates an approximate prediction in the first stage, then crops the image as the input of a second refine network. Nevertheless, this method relies heavily on the accuracy of coarse predictions and does not make good use of the relationship between global and local contexts due to the cropping operation in the coarse stage. In addition, the increase of inference time makes the coarse-to-fine methods less favorable for practical applications. A global scene of the image can provide global semantic information, while local context around the target object produces the relationship between foreground and background. Both of them provides useful hints for inferring the content of the target. Therefore, how to strengthen local context features and retain global context features becomes the key to improve salient object segmentation accuracy. To this end, we propose a novel Local Context Attention Network (LCANet) to adapt the global and the local features with a uniform representational power.

Specifically, the proposed LCANet is mainly built upon a one-stage coarse-to-fine architecture. A coarse segmentation network is built from a standard classification model to extract the discriminant feature, which is up-scaled to generate the coarse feature map. Then an Attentional Correlation Filter (ACF) module is designed to generate local context attention. The local feature map is cropped based on the coarse location after image processing, and is regarded as the correlation filter to make a convolution with the whole feature map. With the help of convolution, the correlation feature map is taken as an attention map to concatenate with the original feature to explicitly enhance local context description ability. In contrast to other attention modules [18, 19, 21] exploring the channel or spatial weight, the proposed ACF module retains the global receptive field with local context enhanced in the surrounding areas, as shown in Figure

2 (d). Furthermore, ACF module is enhanced to Local Context Block (LCB) by Multi-scale ACF operator and a Local Coordinate Convolution (LCC) layer. The LCC layer adopts the relative coordinates of the coarse prediction as another additional feature map to adaptively incorporate the local region context into the global scene context in the spatial dimension. Based on the enhanced LCB, a one-stage coarse-to-fine network is constructed in a type of encoder-decoder architecture as shown in Figure 2. A multi-stage decoder is designed to aggregate the high-level information extracted by coarse network to gradually refine the segmentation results. The LCB is also implemented in the stages of decoder to handle with various sizes of salient objects.

In summary, there are three contributions in our paper:

- We rethink the salient object segmentation task from the intrinsic properties of salient object, and create a newly model structure, Local Context Attention Network (LCANet), to strengthen local context features by adaptively integrating local region context and global scene context in a one-stage coarse-to-fine architecture.
- We design a novel Local Context Block (LCB), on the basis of Attentional Correlation Filter (ACF). As a basic module, it can be used in many situations, instead of the traditional global and non-local attention map.
- Detailed experiments on five widely-used benchmarks indicate the effectiveness of our proposed modules and architecture in improving the accuracy. We achieve the state-of-the-art performance on all of these datasets with thorough ablation studies, especially a new record of **0.883 max F-score** and **0.034 MAE** without any other refinement steps on DUTS-TE dataset.

2   Related Work

Salient Object Segmentation: In the past decades, a number of approaches for saliency detection are developed. Most of them regard salient object segmentation as a special case of semantic segmentation, try to increase receptive field through multi-level features[6, 22, 4, 23–25]. They think high-level features in deep layers encode the semantic information of an abstract description of objects, while low-level features keep spatial details for reconstructing the object boundaries. Some researchers think edge information is the key to improve segmentation accuracy. [22, 5] adopt post-processing heuristics to obtain the refine predictions. DEA[26] simply uses an extra loss to emphasize the detection error for the pixels within the salient object boundaries, while others [27, 28] consider semantic contour information from a pretrained contour detector. Although these methods are proved to be effective, no one analyzes the problem from the perspective of salient objects. Local context is easily overlooked in salient object segmentation.

Attention Module: Attention module is one of the most popular operations in recent neural networks to mimic the visual attention mechanism in the human visual system. SE-Net[18] explores a channel-wise attention map and has
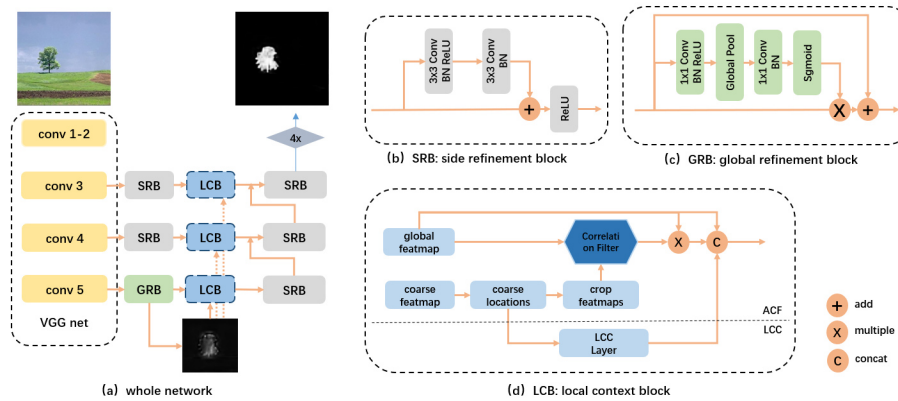
Fig. 2. Overview of our Local Context Attention Network (LCANet). A coarse network is built on a VGG network followed by GRB to generate coarse prediction of the input image. Then the feature map is entered into LCB to produce local context attention maps of the input global features. LCB consists of ACF and LCC modules. It can be implemented on multi-stage of decoder network, cooperating with SRB to gradually refine the final segmentation results.

achieved state-of-the-art performance in image classification. In the field of semantic segmentation, several methods [29–31] adopt multi-scale attention map to increase the receptive field of the high-level features. EncNet[32, 33] introduces context encoding to enhance the prediction that is conditional on the encoded semantics. Non-Local[19, 21] is further proposed self-attention with non-local affinity matrix for vision tasks. By contract, we apply the local context attention map on the output of encoder network to integrate the local region and global scene description.

Correlation Filter: Correlation filter [34–36] has proved to be effective in most of single object tracking methods. It takes advantage of the fact that they can specify the desired output of a linear classifier for several translations and image shifts based on the dot-product at different relative translations of two patches using only a fraction of the computational power. It is relevant to local similarity between the object and its local neighborhood and suitable for convolution operator. Naturally, encoding the local correlation feature as an attention map is an immediate thought.

Coarse-to-fine: These also exist researchers finding the optimal coarse-to-fine solutions to deal with the scale-variation problem. However, most of them focus on the feature expression. DSS [6] proposes a saliency method by introducing short connections to the skip-layer structures within a multi-layer HED architecture. NLDF[37] combines the local and global information through a multi-resolution grid structure. Amulet[4] directly aggregates multi-level features by concatenating feature maps from different layers. Although the features from

deeper layers help to locate the target, objects with different sizes cannot be represented in the same feature structure. A more proper way is to employ the multi-scale features in a coarse-to-fine fashion and gradually predict the final saliency map. However, the increase of inference time makes the coarse-to-fine methods less favorable for practical.

3 Methodology

In this section, we first detailedly introduce the idea of the Attentional Correlation Filter (ACF) module, and elaborate how this module specifically handles the local context issue. After that, we describe the complete Local Context Attention Network (LCANet) based on the Local Context Block (LCB), which builds a coarse-to-fine structure in a one-stage encoder-decoder architecture.

3.1 Attentional Correlation Filter

Let's revisit the main prior of salient object segmentation. Salient object usually has unique feature representations that are different from the surrounding background, such as vivid colors or clear boundaries. Pixels of salient object have strong correlation with its local context in physical space. Based on this observation and analysis, we propose an Attentional Correlation Filter(ACF) module to strengthen the salient features, which apply the attention mechanism to enhance feature expression ability of local context in global feature.

Correlation Filter. Given a target object $T \in \mathbb{R}^{H' \times W'}$ and an reference image $I \in \mathbb{R}^{H \times W}$. For each pixel $(x, y) \in I$, we have the correlation map $Corr$ between I and T is calculated as

$$Corr(I,T)_{x,y} = \sum_{v=-\frac{W'}{2}}^{\frac{W'}{2}} \sum_{u=-\frac{H'}{2}}^{\frac{H'}{2}} T[u,v] \odot I[x+u, y+v] \qquad (1)$$

The correlation map can be generated according to the sliding window on the reference image. In the field of single object tracking[36, 34, 35], it is used widely to obtain the similarity degree between the tracking object and its neighborhood scene to search for the update object position in the next frame. The higher the correlation value, the higher the probability that two patches belong to the same object. The operation of the sliding window is similar to convolution. We regard the feature of the target object as convolution weight, then the correlation feature map can be generated by convolution on the original global feature.

Attentional Correlation Filter. We use the generated correlation feature map as an attention layer for subsequent processing. Given a local feature map $T \in \mathbb{R}^{C' \times H'W'}$ and a global feature map $I \in \mathbb{R}^{C \times HW}$, ACF can be generated as,

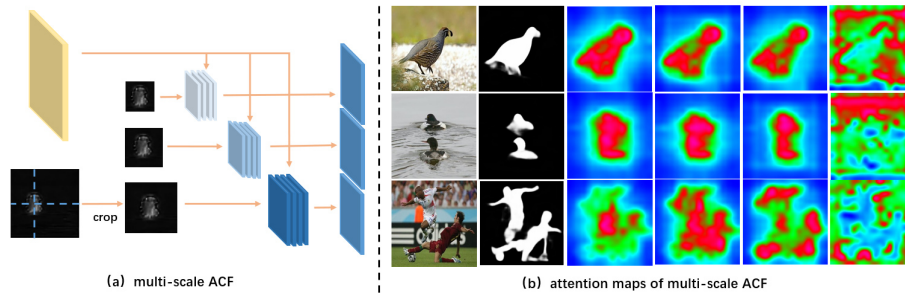$$ACF = I \odot Sigmoid(Corr(I,T)) \qquad (2)$$

Fig. 3. Multi-scale ACF. In (a), multi-scale local feature maps are cropped based on coarse prediction and convoluted with the original feature to generate attention maps. In (b), attention maps of ACF with various scales (line 3,4,5) from coarse results (line 2) are compared with the attention map of non-local module (line 6).

There are many ways to implement attention map, as compared in Figure 1. In the global attention [18] or the non-local affinity attention [19, 21], features are used to calculate the pixel similarity in channel or spatial dimensions. When it acts on the local area, the similar local affinity map can be generated by multiplying the local and global features. However, it is unreasonable that the correlation in channel dimensions loses the spatial relationship between pixels. The direct feature fusion will lead to the migration between features. In ACF, we crop the high-level features from encoder into local feature maps based on the coarse prediction. The coarse object location is calculated from the maximum and the minimum positions of the binarized course predictions. Then the coarse feature map is warped into a given size based on the affine transformation. The generated local context feature map is normalized to keep the same channel of global feature map and odd kernel size, and convoluted with the original feature to generate attention feature map. As shown in Figure 3, the non-local attention map puts more emphasis on bright areas than significant areas, while the ACF significantly enhances the feature discrimination ability in the region of interest.

3.2   Local Context Block

We further extend the proposed ACF module to a more robust local context descriptor. Firstly, we introduce a Local Coordinate Convolution (LCC) layer to increase the local context information in the spatial perspective. Secondly, the ACF is extended to multi-scale ACF to better cope with the significant changes of different sizes of objects, and the inaccurate results of the coarse network.

Local Coordinate Convolution. The calculation of Local Coordinate Convolution Layer lcc_x, lcc_y of target pixel $p_t = I(x_t, y_t) \in I$ can be formulated as following:

$$lcc_x(x, y) = [1 - \frac{|x - x_t|}{H}], \ (x, y) \in H \times W \tag{3}$$

$$lcc_y(x,y) = [1 - \frac{|y - y_t|}{W}], \ (x,y) \in H \times W \qquad (4)$$

we concatenate the LCC and I in channel dimension, and then send them into the convolution layer. The standard coordinate convolution is developed to provide an explicit relative position relationship on the spatial feature map. Based on the coarse location of the salient object, we modify it to further enhance local features in the perspective of physical space. lcc_x and lcc_y are respectively calculated through the distance of pixel relative to the center point of coarse prediction. The nearer the distance, the greater the value. LCC module, combined with ACF module, extracts the local context information of coarse results in the dimensions of feature and space dimensions, in order to complement each other in local area enhancement.

Multi-scale ACF. Furthermore, in order to compensate the error of the coarse prediction, we apply the multi-scale attention maps. Based on the rough segmentation result, the original coarse feature map is warped into different scales. We extract local characteristics with scale changes by adjusting the expansion of the external rectangle of coarse predictions. All these multi-scale attention maps are respectively multipled and concatenated with the global features and to generate the local enhanced feature maps. As shown in Figure 3, with the implementation of Multi-scale ACF, the significance of objects with different sizes is highlighted on different attention maps. On the contrary, the non-local attention more strengthens the distribution of global features, while ignores the salient objects.

3.3 Network Architecture

In the task of semantic segmentation, it has been proved that high-level features and multi-scale decoder are the main factors to improve the segmentation results. While the performance of LCB depends on the result of coarse network, the first stage of LCANet needs to obtain acceptable coarse prediction results under controllable computational complexity. Then a multi-stage decoder is needed to further refine the coarse results. Based on the above knowledge, we propose two complementary modules that are capable of capturing the positions of salient objects and refining their details from coarse level to fine level.

Coarse Network. The most direct way to integrate coarse and refined results is to share the encoder layers. High-level features generated from the encoder network already have preliminary classification ability in terms of distinguishing salient objects and backgrounds. We applied a standard VGG[38] network as the backbone network to extract multi-layer discriminating feature maps. Following the high-level semantic features from VGG, a Global Refinement Block (GRB) is applied to improve receptive field and change the weights of the features to enhance the semantic consistency. As illustrated in Figure 2, two 1x1 convolutions and a global pooling operations are applied to reallocate the feature map and generate a spatial attention map onto the high-level features. This global attention map explicitly makes feature maps be aware of the loca-

tions of the salient objects. Then coarse prediction is generated by up-sampling this global strengthened features.

Coarse-to-fine Network. Based on the coarse network and LCB, we construct a one-stage coarse-to-fine network. We build our network based on a type of classic U-shape architecture as shown in Figure 2. Although GRB improves the receptive filed to capture the global information of the input images, the classification abilities in different stages of backbone network are ignored, which results in diverse consistency manifestation. So we design a residual structure named Side Refinement Block (SRB) to iteratively fuse discriminant features of different scales, which consists of two 3 x 3 convolution operations. Different from previous symmetrical encoder-decoder architectures, we only take the result of stage3 quadrupled as the final output. The last two abandoned up-sampling convolution blocks save a lot of computation, but improve the accuracy in our experiments. As mentioned above, the LCB is implemented cooperating with SRB to strengthen local context features of each stage before up-sampling. Similar to multi scale operations in feature level, multi-stage of LCB is also essential to adapt to the wrong predictions and scale variations as much as possible.

Loss Function. The LCANet is trained in an end-to-end manner with losses of coarse and refine outputs. we use the cross-entropy loss between the final saliency map and the ground truth. Besides that, we apply a boundary preservation loss to enhance the edge accuracy. Sobel operator is adopted to generate the edge of ground truth and the predication saliency map of network. Then the same cross-entropy loss is used to supervise the salient object boundaries. The similar boundary supervision is used and proved in several previous works[1, 39, 23]. Different from them, Sobel operator makes the model pay more attention to the junction of salient objects and background. The total loss function of the LCANet is the weighted sum of the above losses.

$$L = \lambda_0 L_{cs} + \lambda_1 L_{rf} + \lambda_2 L_{cs_{bd}} + \lambda_3 L_{rf_{bd}} \tag{5}$$

Where L_{cs} and L_{rf} denote the cross entropy loss function of the coarse and the coarse-to-fine predictions respectively, and $L_{cs_{bd}}$ and $L_{rf_{bd}}$ are their boundary preservation loss. An experimental weight are applied to combined them all. Also, online hard example mining[40] strategy is adopted with L_{cs} and L_{rf} respectively during training.

In contrast to other two-stage coarse-to-fine networks, LCANet is more like inserting a layer of local context attention into a standard encoder-decoder network. It doesn't separate the coarse and the fine networks with multi times inference. What's more, it adaptively combines the local surrounding context and the global scene information in a uniform feature description.

4   Experiments

In this section, we mainly investigate the effectiveness of the proposed LCB module. While the performance of the LCB depends on the result of coarse network, we first compare the effect of each module in the complete network

structure. Then we conduct more experiments on the configurations of the LCB to analyze the influence of the LCB on local context in detail. Finally, we compare the proposed LCANet with other state-of-the-art approaches quantitatively and qualitatively.

4.1 Experimental Setup

Datasets. To evaluate the performance of the proposed approach, we conduct experiments on five popular benchmark datasets[14, 13, 15–17]. These datasets all contain a large number of images as well as well-segmented annotations and have been widely used in the filed of salient object segmentation. DUTS[14] is the largest dataset containing 10,553 images for training and 5,019 images for testing. Both training and test sets in DUTS contain very complex scenarios with high content variety. ECSSD[13] contains 1,000 natural images manually selected from the Internet. HKU-IS[16] includes 4447 images with multiple disconnected salient objects overlapping the image boundary. DUT-OMRON[15] has 5,168 images with many semantically meaningful and challenging structures. Images of this dataset have one or more salient objects and relatively complex background. PASCAL-S[17] contains 850 natural images that are free-viewed by 8 subjects in eye-tracking tests for salient object annotation.

  Evaluation Metrics. In order to obtain a fair comparison with other state-of-the-art salient object segmentation approaches, we train the proposed networks on DUTS training set (DUTS-TR), and evaluate them on DUTS test set(DUTS-TE) and the other four datasets. For quantitative evaluation, two universally-agreed, standard metrics, mean absolute error (MAE) and maximum F-measure ($maxF$) are adopted respectively [6]. F-measure reflects the overall performance of pixel classification. It is computed by weighted harmonic mean of the precision and recall. MAE indicates the average pixel-wise absolute difference between the estimated salient map and ground-truth.

  Implementation Details. All the networks mentioned below follow the same training strategy. A VGG-16 pre-trained on Imagenet[41] is used to initialize the convolution layers in the backbone network. The parameters in other convolution layers are randomly initialized. All training and test images are resized to 256×256 before being fed into the network. They are trained using mini-batch stochastic gradient descent (SGD) with batch size 48, momentum 0.9, weight decay $1e - 5$ and 300 epochs. As the common configuration, the "poly" learning rate policy is adopted where the initial rate is multiplied by $(1 - \frac{iter}{max\_iter})^{power}$ with power 0.9 and the base learning rate is set as $1e - 4$. Data augmentation contains mean subtraction, random horizontal flip, random resizing with scale ranges in [0.8, 1.2], and random cropping to keep most of the salient object intact for training.

4.2 Ablation Studies of Main Architecture

The main architecture is an encoder-decoder structure consisting of GRB, SRB and LCB. Besides that, boundary loss (BL) is verified here as part of main
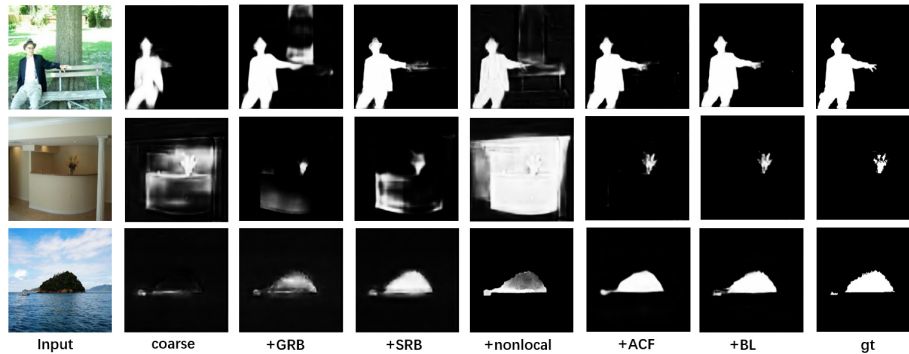
Fig. 4. Result comparisons of different modules in our proposed method. From left to right, each line depicts results of coarse backbone model, and the successive additions of GRB, SRB decoder, non-local attention map, ACF, boundary preservation loss (BL) and the ground truth.

network. We conduct ablation experiments on two challenging datasets DUTS-TE and ECSSD to demonstrate the effectiveness of each module separately.

GRB. Firstly, we verify the accuracy of coarse output. According to Table 1, the addition of GRB gives performance gains in term of both $maxF$ and MAE on the two datasets over the coarse network. While implemented on the coarse network, the $maxF$ of DUTS-TE is improved from 0.775 to 0.818. Also the performance is improved further to 0.843 while applied onto the coarse-to-fine network. The corresponding visual comparisons can be found in Figure 4. The global pooling operator helps to increase the receptive field of the network and highlight the areas of interest.

SRB. Here SRB refers to the decoder network containing the SRB module. As can be observed as in Table 1, simply embedding of SRB decoder helps improve the performance on both $maxF$ and MAE. When GRB and SRB are superposed concurrently, the performance of the model is further enhanced, which indicates the effectiveness of multi-stage refinement for solving the aliasing effect of up-sampling.

ACF. Furthermore, we implement ACF onto the above encoder-decoder network. As the core of LCB module, ACF can well verify the effectiveness of local context attention. Here we only apply the simplest single-scale ACF onto the high-level features generated from coarse network. When it directly acts on the coarse prediction, the accuracy is only slightly improved. This proves from the side that a more reasonable local context is the key to improve the accuracy. When the coarse result is poor, ACF is similar to the attention map on the global context. When it is added to GRB and SRB in turn, the role of ACF is highlighted, and both $maxF$ and MAE on the two datasets are further improved. As shown in Figure 4, the local guidance information generated by ACF

Table 1. Ablation analysis for the main architecture on two popular datasets. All experiments are trained on the union set of DUTS-TR. No.1 depicts the encoder network, and the other modules are successively embedded. As can be observed, each proposed module plays an important role and contributes to the performance.

No.	GRB	SRB	ACF	BL	DUTS-TE		ECSSD	
					$maxF\uparrow$	$MAE\downarrow$	$maxF\uparrow$	$MAE\downarrow$
1					0.775	0.081	0.863	0.080
2	✓				0.818	0.051	0.883	0.066
3		✓			0.830	0.047	0.891	0.058
4	✓	✓			**0.843**	**0.046**	**0.915**	**0.054**
5			✓		0.781	0.077	0.871	0.078
6	✓		✓		0.847	0.044	0.904	0.051
7	✓	✓	✓		0.873	0.039	0.928	0.034
8	✓	✓	✓	✓	**0.875**	**0.038**	**0.931**	**0.033**

allows our network to focus more on the salient objects, and greatly improve the quality of resulting salient maps.

BL. Finally, the boundary preservation loss (BL) is adopted to further improve the quality of boundary in the produced saliency map. Although the performance is only slightly improved, the upper edge of the visual results was optimized to be closer to the ground truth in Figure 4.

4.3   Ablation Studies of LCB

To demonstrate the effectiveness of our proposed LCB, we adopt two basic networks. *coarse* depicts the GRB enhanced coarse network, and *baseline* is the combined model of GRB, SRB and BL, which has been verified in the previous section. We study the different variants of ACF and LCC modules in LCB, and compare them with other attention methods.

Coarse-to-fine Structures. By introducing local context into the baseline network, the most direct way is to execute the model twice by cropping the predictions of the first network as the input of the second one. In Table 2, we conducted the experiments of executing the *coarse* and *baseline* twice respectively. The performance of $baseline \times 2$ is significantly improved both in terms of $maxF$ and MAE, while the accuracy of $coarse \times 2$ is slightly decreased. This is mainly because the refine network is heavily dependent on the results of coarse network. However, when $baseline + ACF$ integrates the local context attention and the global feature map in a joint training manner, the performance is greatly improved as shown in Table 2.

Other Attention Modules. Also we compare the proposed ACF module with other attention-based modules. Att_{se}, Att_{nl} and Att_{lf} are effective modules of squeeze-and-excitiation[18], non-local affinity[19], and a modified local affinity depicted in Figure 1. As can be observed from Table 2, although the accuracy can be improved by increasing the receptive field, these modules are still

Table 2. Ablation analysis for the proposed main architecture on DUTS-TE. Different types of ACF and LCC modules are verified and compared with other attention modules. As can be noticed, ACF is superior to others, and LCC also contributes to the performance.

Model	$maxF\uparrow$	$MAE\downarrow$	Model	$maxF\uparrow$	$MAE\downarrow$
coarse $\times$ 2	0.813	0.050	baseline + $ACF_{0.1}$	0.874	0.038
baseline $\times$ 2	0.861	0.042	baseline + $ACF_{0.3}$	0.871	0.041
baseline + Att_{se}	0.845	0.047	baseline + $ACF_{0.5}$	0.875	0.038
baseline + Att_{lf}	0.847	0.046	baseline + ACF_{msz}	0.879	0.037
baseline + Att_{nl}	0.861	0.041	baseline + ACF_{msz} + LCC	0.882	0.041
baseline + ACF	**0.875**	**0.038**	baseline + $ACF_{msz+msa}$ + LCC	**0.883**	**0.034**

worse than the proposed ACF module. Att_{lf} is similar to ACF by generating the affinity matrix bewteen local context and global feature map. Nevertheless, the performance is even worse than results of $baseline \times 2$, we think that the multiplication of affinity matrix in different feature space destroys the ability of feature description of both global and local feature map.

Multi-scale ACF. Furthermore, we investigates the effectiveness of Multi-scale ACF. Figure 3 depicts the visualization results of attention maps produces by the cropped feature maps with different sizes. Also the performance of Multi-scale ACF are showed in Table 2. It can be observed both in qualitative or quantitative analysis that different attention map corresponds to different size of salient object. And Also the unknown segmentation errors of coarse network also affect the performance of ACF. Although the accuracy of ACF with different scales fluctuates, Multi-scale ACF can cope with these accidental changes, and further improve the performance of both $maxF$ and MAE.

LCC. LCC is another complementary module in LCB. As present in Table 2, LCC further improves the result of Multi-scale ACF from 0.879 to 0.882. This proves the complementary relationship between ACF and LCC. Under the guidance of an explicit location provided by LCC, it realizes a more accurate and robust segmentation on the final result. The final LCB consists of Multi-scale ACF and LCC. It can be implemented onto the different stages of decoder to further enhance the final performance as $ACF_{msz+msa}$ depicted in Table 2.

4.4 Comparisons with the State-of-the-Arts

We compare the proposed LCANet with sixteen recent state-of-the-art methods including PoolNet[1], PFA[39], AFNet[2], MLMSNet[3], CPD[42], BDMPM[23], GRL[43], PAGRN[44], Amulet[4], SRM[45], UCF[46], DCL[5], DHS[47], DSS[6], ELD[24] and NLDF[37] on five datasets. We obtain the saliency maps of these methods from authors or the deployment codes provided by authors for fair comparison.

Quantitative Evaluation. Table 3 shows the quantitative evaluation results of the proposed method and other state-of-the-art salient segmentation approaches in terms of $maxF$ and MAE. As present, our method outperforms other

Table 3. Quantitative comparisons of the proposed approach and sixteen state-of-the-art CNN based salient object detection approaches on five datasets. The best two scores are shown in red and blue.

	DUTS-TE		ECSSD		HKU-IS		PASCAL-S		DUT-OM	
	$maxF$	MAE	$maxF$	MAE	$maxF$	MAE	$maxF$	MAE	$maxF$	MAE
ELD[24]	0.737	0.092	0.867	0.081	0.840	0.073	0.788	0.122	0.719	0.090
DHS[47]	0.811	0.065	0.904	0.062	0.890	0.053	0.845	0.096	-	-
DCL[5]	0.785	0.081	0.895	0.079	0.889	0.063	0.845	0.111	0.756	0.086
UCF[46]	0.772	0.111	0.901	0.070	0.887	0.062	0.849	0.109	0.729	0.120
SRM[45]	0.826	0.058	0.915	0.056	0.905	0.046	0.867	0.085	0.769	0.069
Amulet[4]	0.777	0.084	0.913	0.060	0.896	0.051	0.861	0.098	0.742	0.097
NLDF[37]	0.812	0.064	0.903	0.065	0.901	0.048	0.851	0.100	0.753	0.079
DSS[6]	0.813	0.064	0.895	0.064	0.901	0.047	0.850	0.099	0.760	0.075
PAGRN[44]	0.854	0.054	0.923	0.064	0.917	0.047	0.869	0.094	0.770	0.070
GRL[43]	0.834	0.050	0.923	0.044	0.913	0.037	0.881	0.079	0.778	0.063
BDMPM[23]	0.851	0.048	0.925	0.048	0.920	0.039	0.880	0.078	0.774	0.064
CPD[42]	0.864	0.043	0.936	0.040	0.924	0.033	0.866	0.074	0.794	0.057
MLMSNet[3]	0.851	0.049	0.928	0.045	0.921	0.039	0.862	0.074	0.774	0.064
AFNet[2]	0.862	0.046	0.935	0.042	0.923	0.036	0.868	0.071	0.797	0.057
PFA[39]	0.870	**0.040**	0.931	**0.032**	0.926	**0.032**	**0.892**	**0.067**	**0.855**	**0.041**
PoolNet[1]	**0.880**	0.041	**0.937**	0.044	**0.931**	0.033	0.865	0.072	0.821	0.056
LCANet	**0.883**	**0.034**	**0.939**	**0.029**	**0.931**	**0.030**	**0.889**	**0.064**	**0.843**	**0.037**

approaches across all these datasets. To be specific, LCANet achieves large improvement compared with the best existing approach on DUT-TE dataset. Both of $maxF$ and MAE are definitely increased from the ever best PoolNet[1] based on the VGG backbone. On PASCAL-S and DUT-OMRON, although the performance of $maxF$ is slightly lower than PFA[39], the MAE exceeds it. We find that objects in them are large or multi-subjects. When the object size is large, local context is almost the same as global context the, the role of LCANet will be correspondingly weakened. While the proposed approach is a simple one-stage structure, it can be further improved with multi-branch learning algorithms.

Qualitative Evaluation. To further explain the advantages of our approach, Figure 5 provides a visual comparison of our method and other state-of-the-arts. From the former 5 rows of Figure 5, it is clear that our method is obviously superior to others coping with small objects. While other methods are difficult to distinguish salient objects and background, LCANet obtains accurate segmentation results based on the coarse-to-fine guidance. This further verifies the effectiveness of local context in salient object location. More than this, we notice that the consistency of larger objects is also preserved as shown in row 6,7 of Figure 5. Even when salient objects are scattered in different places of the image, there is still a certain probability not be affected. Also it still obtains a good segmentation prediction when the salient object located in the side of the image. These observations indicate the intergration of local and global context
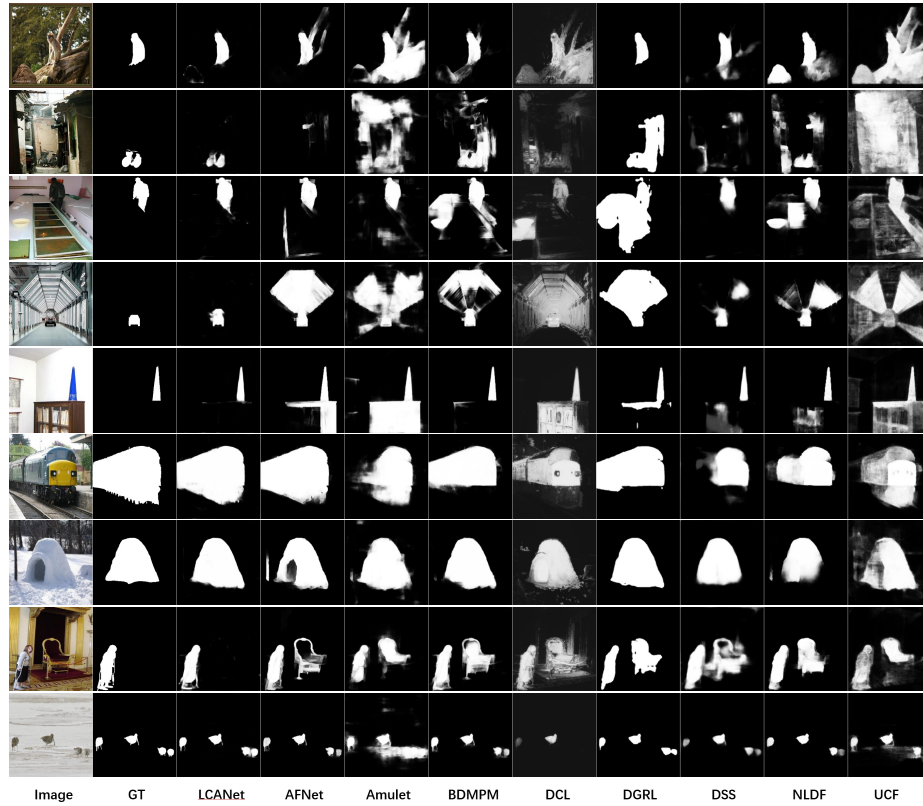
Fig. 5. Qualitative comparisons of the state-of-the-art algorithms and our proposed **LCANet**. GT means ground-truth masks of salient objects.

information is important to deal with salient object segmentation, regardless of the position and size of the object.

5   Conclusion

In this paper, we propose a local context attention network to cope with salient object segmentation. Based on the prior that saliency object usually has unique feature representations that are different from the surrounding background, we proposed a Local Context Block consisting of an Attentional Correlation Filter and a Local Coordinate Convolution layer, in order to intergrate the local context information into the global scene features in a one-stage coarse-to-fine architecture. Detailed experiments verify the feasibility of the proposed LCANet. It achieves comparable performance with other state-of-the-art methods based on a simple baseline. We believe that this model will be useful for other scenarios.

References

1. Liu, J., Hou, Q., Cheng, M.M., Feng, J., Jiang, J.: A simple pooling-based design for real-time salient object detection. In: CVPR. (2019) 3917–3926
2. Feng, M., Lu, H., Ding, E.: Attentive feedback network for boundary-aware salient object detection. In: CVPR. (2019) 1623–1632
3. Wu, R., Feng, M., Guan, W., Wang, D., Lu, H., Ding, E.: A mutual learning method for salient object detection with intertwined multi-supervision. In: CVPR. (2019) 8150–8159
4. Zhang, P., Wang, D., Lu, H., Wang, H., Ruan, X.: Amulet: Aggregating multi-level convolutional features for salient object detection. In: ICCV. (2017) 202–211
5. Li, G., Yu, Y.: Deep contrast learning for salient object detection. In: CVPR. (2016) 478–487
6. Hou, Q., Cheng, M.M., Hu, X., Borji, A., Tu, Z., Torr, P.H.S.: Deeply supervised salient object detection with short connections. In: CVPR. (2017) 5300–5309
7. Liu, Y., Zhang, Q., Zhang, D., Han, J.: Employing deep part-object relationships for salient object detection. In: ICCV. (2019) 1232–1241
8. Wang, W., Zhao, S., Shen, J., Hoi, S.C.H., Borji, A.: Salient object detection with pyramid attention and salient edges. In: CVPR. (2019) 1448–1457
9. Zeng, Y., Zhang, P., Zhang, J., Lin, Z.L., Lu, H.: Towards high-resolution salient object detection. In: ICCV. (2019)
10. Zeng, Y., Zhuge, Y.Z., Lu, H., Zhang, L., Qian, M., Yu, Y.: Multi-source weak supervision for saliency detection. In: CVPR. (2019) 6074–6083
11. Zhang, L., Zhang, J., Lin, Z., Lu, H., He, Y.: Capsal: Leveraging captioning to boost semantics for salient object detection. In: CVPR. (2019) 6024–6033
12. Zhao, J., Liu, J., Fan, D.P., Cao, Y., Yang, J., Cheng, M.M.: Egnet: Edge guidance network for salient object detection. In: ICCV. (2019)
13. Yan, Q., Xu, L., Shi, J., Jia, J.: Hierarchical saliency detection. In: CVPR. (2013) 1155–1162
14. Wang, L., Lu, H., Wang, Y., Feng, M., Wang, D., Yin, B., Ruan, X.: Learning to detect salient objects with image-level supervision. In: CVPR. (2017) 3796–3805
15. Yang, C., Zhang, L., Lu, H., Ruan, X., Yang, M.H.: Saliency detection via graph-based manifold ranking. In: CVPR. (2013) 3166–3173
16. Li, G., Yu, Y.: Visual saliency based on multiscale deep features. In: CVPR. (2015) 5455–5463
17. Li, Y., Hou, X., Koch, C., Rehg, J.M., Yuille, A.L.: The secrets of salient object segmentation. In: CVPR. (2014) 280–287
18. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: CVPR. (2018) 7132–7141
19. Wang, X., Girshick, R.B., Gupta, A., He, K.: Non-local neural networks. In: CVPR. (2018) 7794–7803
20. Piao, Y., Ji, W., Li, J., Zhang, M., Lu, H.: Depth-induced multi-scale recurrent attention network for saliency detection. In: ICCV. (2019)
21. Yuan, Y., Wang, J.: Ocnet: Object context network for scene parsing. (2019)
22. Wang, L., Wang, L., Lu, H., Zhang, P., Ruan, X.: Saliency detection with recurrent fully convolutional networks. In: ECCV. (2016) 825–841
23. Zhang, L., Dai, J., Lu, H., He, Y., Wang, G.: A bi-directional message passing model for salient object detection. In: CVPR. (2018)
24. Lee, G., Tai, Y.W., Kim, J.: Deep saliency with encoded low level distance map and high level features. In: CVPR. (2016) 660–668

25. Tang, Y., Wu, X., Bu, W.: Deeply-supervised recurrent convolutional neural network for saliency detection. In: ACM Multimedia. (2016) 397–401
26. Zhang, J., Dai, Y., Porikli, F., He, M.: Deep edge-aware saliency detection. (2017)
27. Li, X., Yang, F., Cheng, H., Liu, W., Shen, D.: Contour knowledge transfer for salient object detection. In: ECCV. (2018) 370–385
28. Yang, J., Price, B.L., Cohen, S., Lee, H., Yang, M.H.: Object contour detection with a fully convolutional encoder-decoder network. In: CVPR. (2016) 193–202
29. Zhao, H., Shi, J., Qi, X., Wang, X., Jia, J.: Pyramid scene parsing network. In: CVPR. (2017) 6230–6239
30. Chen, L.C., Zhu, Y., Papandreou, G., Schroff, F., Adam, H.: Encoder-decoder with atrous separable convolution for semantic image segmentation. In: ECCV. (2018) 833–851
31. Li, H., Xiong, P., An, J., Wang, L.: Pyramid attention network for semantic segmentation. In: BMVC. (2018)
32. Zhang, H., Dana, K.J., Shi, J., Zhang, Z., Wang, X., Tyagi, A., Agrawal, A.: Context encoding for semantic segmentation. In: CVPR. (2018) 7151–7160
33. Chen, L.C., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: Scale-aware semantic image segmentation. In: CVPR. (2016) 3640–3649
34. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.P.: High-speed tracking with kernelized correlation filters. (2015) 583–596
35. Henriques, J.F., Caseiro, R., Martins, P., Batista, J.P.: Exploiting the circulant structure of tracking-by-detection with kernels. In: ECCV. (2012) 702–715
36. Bolme, D.S., Beveridge, J.R., Draper, B.A., Lui, Y.M.: Visual object tracking using adaptive correlation filters. In: CVPR. (2010) 2544–2550
37. Luo, Z., Mishra, A.K., Achkar, A., Eichel, J.A., Li, S., Jodoin, P.M.: Non-local deep features for salient object detection. In: CVPR. (2017) 6593–6601
38. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: ICLR. (2015)
39. Zhao, T., Wu, X.: Pyramid feature attention network for saliency detection. In: CVPR. (2019) 3085–3094
40. Shrivastava, A., Gupta, A., Girshick, R.B.: Training region-based object detectors with online hard example mining. In: CVPR. (2016) 761–769
41. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: NIPS. (2012)
42. Wu, Z., Su, L., Huang, Q.: Cascaded partial decoder for fast and accurate salient object detection. In: CVPR. (2019) 3907–3916
43. Wang, T., Zhang, L., Wang, S., Lu, H., Yang, G., Ruan, X., Borji, A.: Detect globally, refine locally: A novel approach to saliency detection. In: CVPR. (2018)
44. Zhang, X., Wang, T., Qi, J., Lu, H., Wang, G.: Progressive attention guided recurrent network for salient object detection. In: CVPR. (2018) 714–722
45. Wang, T., Borji, A., Zhang, L., Zhang, P., Lu, H.: A stagewise refinement model for detecting salient objects in images. In: ICCV. (2017)
46. Zhang, P., Wang, D., Lu, H., Wang, H., Yin, B.: Learning uncertain convolutional features for accurate saliency detection. In: ICCV. (2017) 212–221
47. Liu, N., Han, J.: Deep hierarchical saliency network for salient object detection. In: CVPR. (2016) 678–686