Multi-scale fusion and non-local attention mechanism based salient object detection

Liyuan Chen[1], *Zhonglong* Zheng[1,*], *Pengcheng* Bian[1], *Jiashuaizi* Mo[1], and *Abd Erraouf* Khodja[1]

[1]Zhejiang Normal University, Jinhua, Zhejiang, China

Keywords: Salient object detection, Multi-scale fusion, Guidance.

Abstract. With the development of deep learning, researches in the field of computer vision are attracting more attention. As the pre-processing operation of visual tasks, a salient model may focus on pure architectures. The paper proposes a new multi-scale fusion network to enrich high-level redundant information with the enlarged receptive field. With the guidance of attention mechanism, the framework can capture more effective correlation spatial and channels information. Building a short-connection between high-level and each level features to transmit the contextual features. The model can be used in a variety of complex scenes for end-to-end image detection, with simple structure and strong versatility. Experimental results obtained on multiple common datasets have shown that the proposed model achieved better performance both in the visual effect and the accuracy for small object and multi-target detection.

1 Introduction

Salient object detection aims to identify the distinctive visually objects or regions in images then distinguish them from the environments [1]. While the conventional convolutional neural networks have obtained excellent performances in the task of salient object detection, there still exist some problems resulting in it suboptimal.

During the process of feature extraction through the convolutional neural networks, the image resolution reduced after repeated pooling operation generally. For better semantic salient object detection, high-level features matter, however, the quality of the prediction cannot be guaranteed by using up-sampling.

For an end-to-end model, the prediction based on the attention module performs better. Based on the result of dilation multi-scale fusion, the proposed attention module refined non-local network, and the spatial and channel separately of each layer can be enhanced, similar with ResNet. Multi-scale fusion is to enhance the spatial features without flexible-shape images that are pre-processed. Meanwhile, more effective features based on the separate spatial and channel can be extracted using non-local attention mechanism.

In this paper, we purpose a novel salient object detection deep neural network, multi-scale fusion based on non-local attention with short connection network (FNASNet). The previous

* Corresponding author: 479059697@qq.com

works showed that the multi-scale fusion mechanism and the attention mechanism make great contributions to extracting flexible features. FNASNet adopts the multi-scale fusion mechanism with dilation for more complete extraction of global and local features. Besides, non-local attention module infers global features along two separate dimensions, spatial and channel, making the network concentrate on more effective information. Short connections enable the transfer of feature from high-level to low-level, which facilitates the network to capture context.

2 The proposed architecture

2.1 Network overview

Based on VGG-16 Net, the multi-scale fusion module aims at extracting multi-scale contextual features of images with the guidance of non-local attention module, which effectively selects efficient features on spatial and channel dimension.See Figure 1.

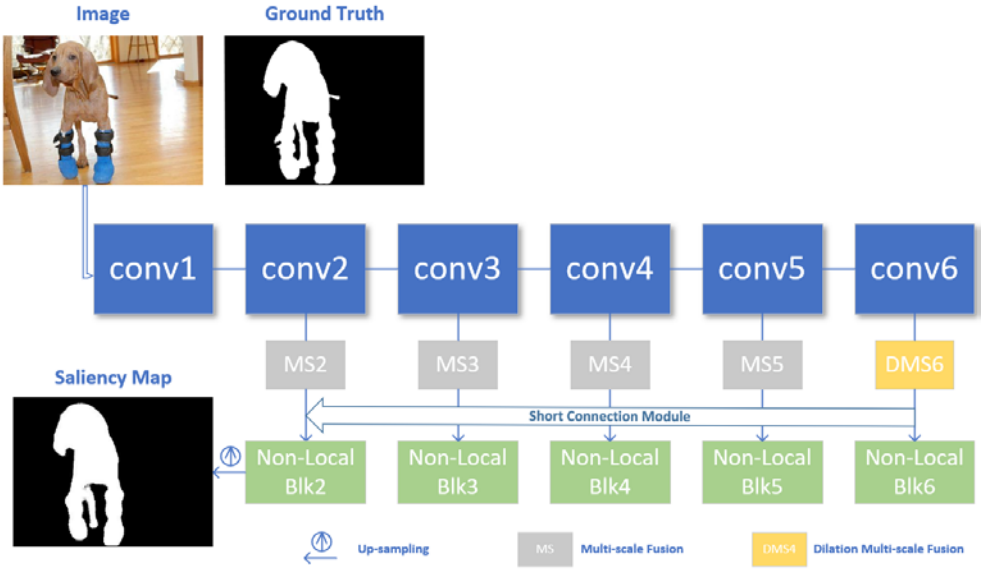

Fig. 1. Network architecture.

2.2 Multi-scale fusion module

This network adopts multi-scale fusion within the convolutional block by using 1×1 kernel to fuse the multi-scale features after dilation. Multi-scale features of an image can be cascaded together using different kernels. Figure 2 shows the multi-scale fusion module adopted in this paper.

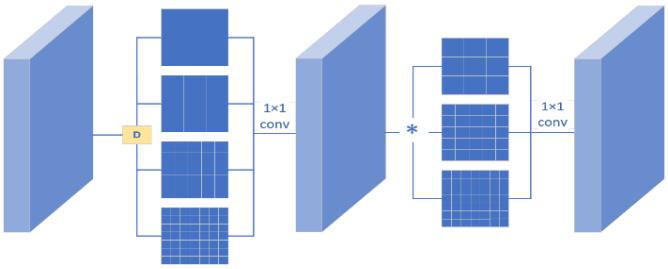

Fig. 2. Multi-scale fusion module.

Inspired by the dilation in morphology[2], the network increases the redundancy by using dilation strategy to make up for the loss of high-level features, on the basis of the extracted high-level features, and avoid introducing too much redundant information to affect the prediction. Morphological dilation $D^l(\cdot)$ is implemented by utilizing the specific convolution with multi-scale, written as,

$$D^{(l)} = f^{(l)} * W_d^{(l)} \tag{1}$$

where $f^{(l)}$ represents the l^{th} feature map, $W_d^{(l)}$ is the dilation kernel of l^{th} layer, $l = 6$, the dilation rate rate is set to size of 1,3,5,7.

$f_{ms}^{(l)}$ is used to represent the multi-scale fusion features, which can be obtained from equation (2):

$$f_{ms}^{(l)} = M^{(l)} * W_{ms}^{(l)} \tag{2}$$

$$F_{ms}^{(l)} = \sigma\left(f_{ms}^{(l)}\right) \tag{3}$$

where M^l represents the morphology operation. When $l = 6$, this operation act as dilation. $W_{ms}^{(l)}$ represents a multi-scale fusion kernel.

2.3 Non-local attention module

All the pixel position and channel can be calculated with the softmax operation from feature map, aiming to give effective guidance for assigning global contextual attention on each pixel. Figure 3 shows the non-local spatial and channel attention module.

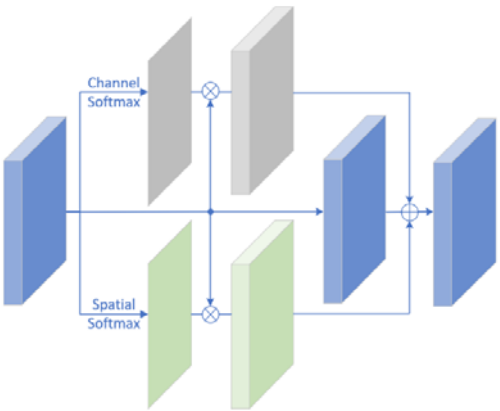

Fig. 3. Non-local attention block.

Non-local attention module aims at generating an attention map at each pixel and channels over its context region and constructing an contextual feature with attention to enhance the feature representability of the network.

At l^{th} feature map, the obtained feature vector about spatial and channels, which are denoted as $\boldsymbol{x}_{w,h}$ and $\boldsymbol{x}_c, \boldsymbol{x}_{w,h} \in R^{W\times H}, \boldsymbol{x}_c \in R^C$, are calculated via a softmax function to generate the global attention weights $\boldsymbol{\alpha}_{w,h}$ and $\boldsymbol{\alpha}_c$ from spatial and channel, respectively.

$$\alpha_{w,h}^i = \frac{\exp\left(x_{w,h}^i\right)}{\sum_{i=1}^{C}\exp\left(x_{w,h}^i\right)} \tag{4}$$

$$\alpha_c^i = \frac{\exp(x_c^i)}{\sum_{i=1}^{C}\exp(x_c^i)} \tag{5}$$

where i represents the number of channels for each pixel (w,h), $\boldsymbol{\alpha}_{w,h} \in R^{W\times H}, \boldsymbol{\alpha}_c \in R^C$.

For the pixel (w,h), the features at all locations in $\boldsymbol{f}_{w,h}^l$ are weighted by $\boldsymbol{\alpha}_{w,h}$ to construct the attended spatial contextual feature $\boldsymbol{f}_{w,h}^l$,

$$\boldsymbol{f}_{w,h}^l = \alpha_{w,h}^l * x_{w,h}^l \tag{6}$$

For the channel c, the features at all channels in $\boldsymbol{f}^c$ are weighted by $\boldsymbol{\alpha}_c$ to construct the attended channel contextual feature,

$$\boldsymbol{f}_c^l = \alpha_c^l * x_c^l \tag{7}$$

Finally, the non-local attention at each spatial and channel feature F_{nla}^l is the sum of $\boldsymbol{f}_{w,h}^l$, $\boldsymbol{f}_c^l$ *and* $\boldsymbol{x}_{w,h}^l$,

$$F_{nla}^l = \sum_{l=1}^{L}\boldsymbol{f}_{w,h}^l + \boldsymbol{f}_c^l + \boldsymbol{x}_{w,h}^l \tag{8}$$

2.4 Short connection method

Multi-scale responses are learned from different layers with increasingly larger receptive fields, and these responses are concatenated together for outputting final saliency. To obtain more information, we use short connection transmit the features from high-level to low-level. Figure 4 shows the architecture of short connection module.

Features are first generated via multi-scale fusion based on dilation, merging the information from deep layer to shallow layer to enhance the power of the network.

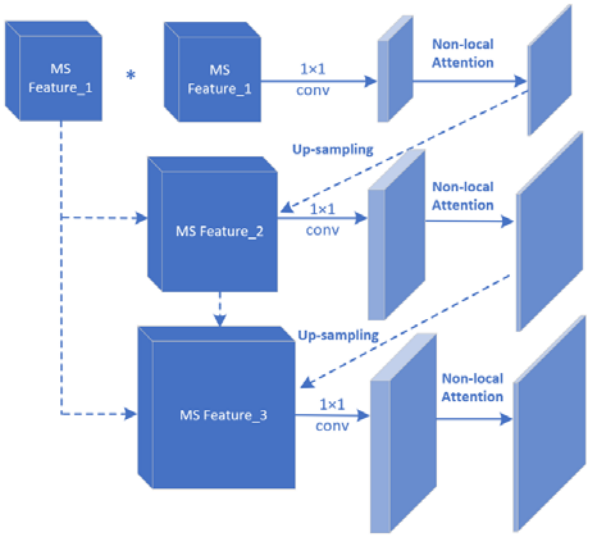

Fig. 4. Short connection module

3 Experiments

3.1 Datasets

In the experiment in this paper, DUTS[3] dataset was used in the training stage, which is the largest public dataset for salient object detection at present. Besides, this method is evaluated in five benchmark datasets,, ECSSD[4], PASCAL-S [5], DUT-O [6], HKU-IS [7] and SOD[8].

3.2 Evaluation metrics

The detection performance of the model in this paper was evaluated by constructing a calculating Mean Absolute Error(MAE) and F-measure. F-measure. It is an overall performance measurement and computed considers from precision and Recall values, as follows:

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

Mean Absolute Error. To analyze the similar between salient map and the ground truth, indicate the impact of non-salient pixels, which is given by the following equation:

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

where $G(x,y)$ represents the ground truth value at the pixel point (x,y), $S(x,y)$ is the salient map.

3.4 Comparisons with the State-of-the-Arts

In this section, we compare our FNAsNet with previous state-of-the-art methods, including MCDL[9], DS[10], DLS[11], SBF[12] and RSDNet[13]. The result shown in Table 1.

Table 1. Different saliency detection methods on six large-scale saliency detection datasets.

	DUT		**ECSSD**		**HKU**		**SOD**		**DUTS**		**PAS**	
	MAE	maxF	MAE	maxF	MAE	maxF	MAE	MaxF	MAE	maxF	MAE	maxF
MCDL[15]	0.089	0.670	0.101	0.816	0.092	0.787	0.182	0.689	0.105	0.634	0.143	0.706
DS[16]	0.120	0.708	0.122	0.868	0.078	0.848	0.190	0.757	0.090	0.747	0.175	0.718
DLS[17]	**0**.090	0.644	0.086	0.826	0.069	0.807	-	-	-	-	0.130	0.712
SBF[17]	0.110	0.649	0.091	0.833	0.078	0.821	0.160	0.740	0.109	0.657	0.133	0.726
RSDNet[18]	0.178	**0.715**	0.173	**0.880**	0.156	**0.871**	0.226	**0.790**	0.161	**0.798**	**-**	**-**
FNAsNet	**0.087**	0.697	**0.085**	0.863	**0.062**	0.848	**0.158**	0.765	**0.082**	0.735	**0.126**	**0.773**

Bold font represents the optimal effect on the dataset, - indicates that the result is not available, the index of the algorithm indicates the year of publication.

Quantitative Evaluation. As shown in Table 1, FNAsNet constructs a new global multi-scale fusion module, most of the comparison algorithms involved in the experiment are based on multi-scale fusion methods. Compared to the previous algorithm, our algorithm have a better performance, the model focus on more effective information, making the relationship between pixels closer, F-measure also is high, in the case of small amplitude fluctuation precision error, improving the detection performance of the model. The model shows good performance on PAS and SOD, and strong ability for small datasets of multi-target detection. The MCDL, DLS, SBF and RSDNet algorithm in the 1st row, 3rd row, 4th row and 5th row, respectivily, all propose the algorithm for fusion of context information and different level feature maps. Our algorithm uses a short connection to cascade high-level features with each level of features, enhancing the context information of the image and achieving good performance. More than one network is constructed in the second of the model, and as the results showing, our method has achieved a better effect than DS algorithm.

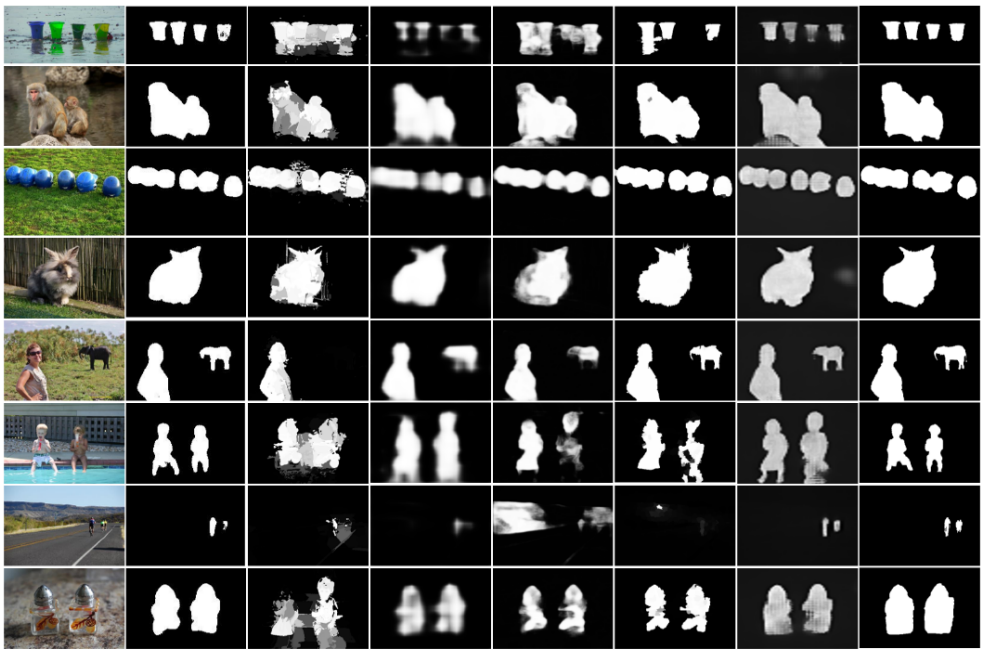

Fig. 5. Visual comparison with state-of-the-art methods.

Qualitative Evaluation. Figure 5 illustrates the visual comparison of our method with other approachs. The first row shows the reflection problem in the image. The results shows that MDCL and DLS have poor performance for reflection problems, and DS did not retain enough integrity of the target. SBF and the algorithm presented in this paper show good results. The second row is the case that the image is similar to the background color. Except for DLS and SBF algorithm, the detection results of other methods are relatively accurate. 3rd, 4th, and 5th rows respectively show the detection performance of different algorithms on multi-target, single-target, and multi-category targets. It is worth noting that the detection results of the last three rows are figure in the reflection, small object detection and transparent object detection respectively. Our algorithm have strong detection performance for small object detection and reflection problem, but on the transparent object detection, target integrity reduce to a certain extent, while other algorithms have a certain limitation on different situations, generally get poor performance on the reflection problem.

4 Summary

In order to enhance the ability of network to extract effective features, this paper proposes a new multi-scale fusion strategy, which uses dilation operation to increase the redundant information of high-level feature maps. Under the guidance of the attention mechanism between spaces and channels, the feature context information is effectively increased through the cascade of short connections between different multi-scale features. Through quantitative experiments and qualitative analysis, the algorithm presented in this paper have high robustness and accuracy. It can be seen from the detection results that the algorithm shows a better detection performance on complex scenes, single object detection and reflection. The experimental results shows that the multi-scale fusion algorithm based on attention mechanism has high credibility and value for constructing an end-to-end salient object detection algorithm at arbitrary scale. In the future, on the basis of this work, more attention

will be paid to the detailed features, such as the edge of the image, to achieve a greater breakthrough in the performance of the model.

This research was financially supported by the No. 61672467, No.61976195 and No.11871438 National Science Foundation.

References

1. L. Itti. Automatic foveation for video compression using a neurobiological model of visual attention. IEEE Trans. Image Process, 1304-1318(2014).
2. M. Feng, H. Lu, E. Ding. Attentive Feedback Network for Boundary-Aware Salient Object Detection. Internaltional Conference on Computer Vision and Pattern Recogintion 1623-1632(2019).
3. L. Wang, H. Lu, Y. Wang, M. Feng, D. Wang, B. Yin, X. Ruan. Learning to detect salient objects with image-level supervision. Internaltional Conference on Computer Vision and Pattern Recogintion, 3796-3805(2017).
4. Q. Yan, L. Xu, J. Shi, J. Jia. Hierarchical saliency detection. Internaltional Conference on Computer Vision and Pattern Recogintion, 1155-1162(2013).
5. Y. Li, X. Hou, C. Koch, J. M. Rehg, A. L. Yuille. The secrets of salient object segmentation. Internaltional Conference on Computer Vision and Pattern Recogintion, 280-287(2014).
6. C. Yang, L. Zhang, H. Lu, X. Ruan, M.-H. Yang. Saliency detection via graph-based manifold ranking. Internaltional Conference on Computer Vision and Pattern Recogintion, 3166-3173(2013).
7. G. Li, Y. Yu. Visual saliency based on multiscale deep features. Internaltional Conference on Computer Vision and Pattern Recogintion, 5455-5463(2015).
8. V. Movahedi and J. H. Elder, "Design and perceptual validation of performance measures for salient object segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops, 2010.
9. R. Zhao, W. Ouyang, H. Li, X. Wang. Saliency detection by multi-context deep learning. Internaltional Conference on Computer Vision and Pattern Recogintion, 1265-1274(2015).
10. X. Li, L. Zhao, L. Wei, M.-H. Yang, F. Wu, Y. Zhuang, H. Ling, and J. Wang, "Deepsaliency: Multi-task deep neural network model for salient object detection," IEEE Trans. Image Process., vol. 25,no. 8, pp. 3919 – 3930(2016).
11. P. Hu, B. Shuai, J. Liu, G. Wang. Deep level sets for salient object detection. Internaltional Conference on Computer Vision and Pattern Recogintion, 540-549(2017).
12. D. Zhang, J. Han, and Y. Zhang, "Supervision by fusion: Towards unsupervised learning of deep salient object detector," in Proc. IEEE Int. Conf. Comput. Vis., vol. 1, no. 2, 2017, p. 3.
13. M. Amirul Islam, M. Kalash, and N. D. B. Bruce, "Revisiting salient object detection: Simultaneous detection, ranking, and subitizing of multiple salient objects," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit.(2018).